Implications of Bioinformatics Applications for High Performance Computing Architectures

July 25, 2012

Bill Feiereisen

Intel Corporation, University of New Mexico

Purpose

- With the advent of next generation sequencing equipment the amount of genomic and proteomic data has exploded. This is not news.
- It has been anticipated for at least 10 years, and has engaged the attention of life scientists who have not traditionally been taught statistics, math and computer science.
- It has also engaged the attention of computer and computational scientists from the traditional high performance computing community, who seek to support life sciences with the tools they have developed for physics.
- Some of these tools are directly applicable – but some are not.
- The matching of life sciences computing needs with the traditional HPC community offerings have not always been smooth.
- However … some of the most exciting scientific work occurs at the boundary between two fields of study. This conjunction of life sciences and HPC is just that.
- This lecture will explore that intersection.

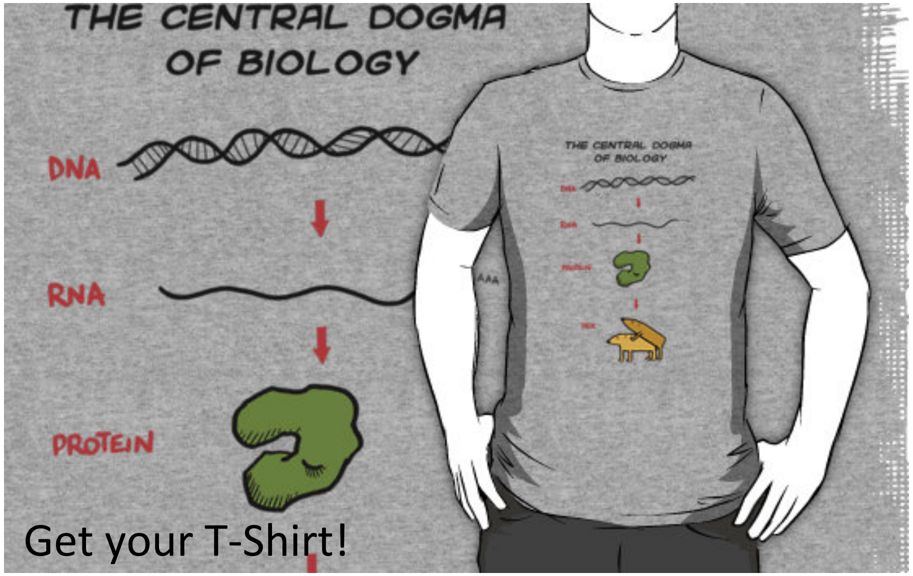
Caveats

- Keep in mind that I am not teaching a bioinformatics course, but using some sampling of bioinformatics algorithms to illuminate the underlying compute machinery and computing needs.
- This is a learning process for all of us (me especially) from multiple disciplines. I'm actually a mechanical engineer, but have worked as computational scientist for my whole career. I am not a bio-informatician
- But I will soon be able to play one on TV!
- As I constructed this lecture I presumed that I would be able to categorize everything into a neat taxonomy. I can't – It's actually the Wild West out there right now. But that is SO exciting.
- Please help me with questions. Let's make this course into a discussion where we both can learn.

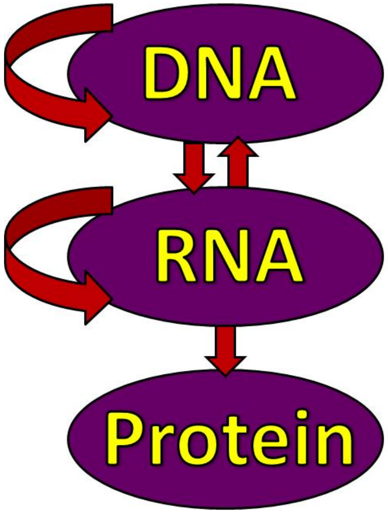
Outline

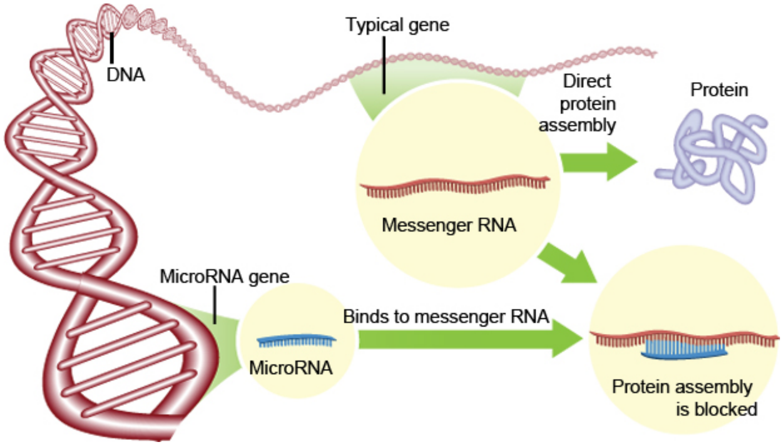
- Some suggestive workflows and the applications they require
  - Biased toward tasks that require high throughput sequencing … and thus require HPC
- An overview of the common applications used in the workflows (and many more) and a description of their algorithms
- The computing needs of these algorithms

The central dogma of biology



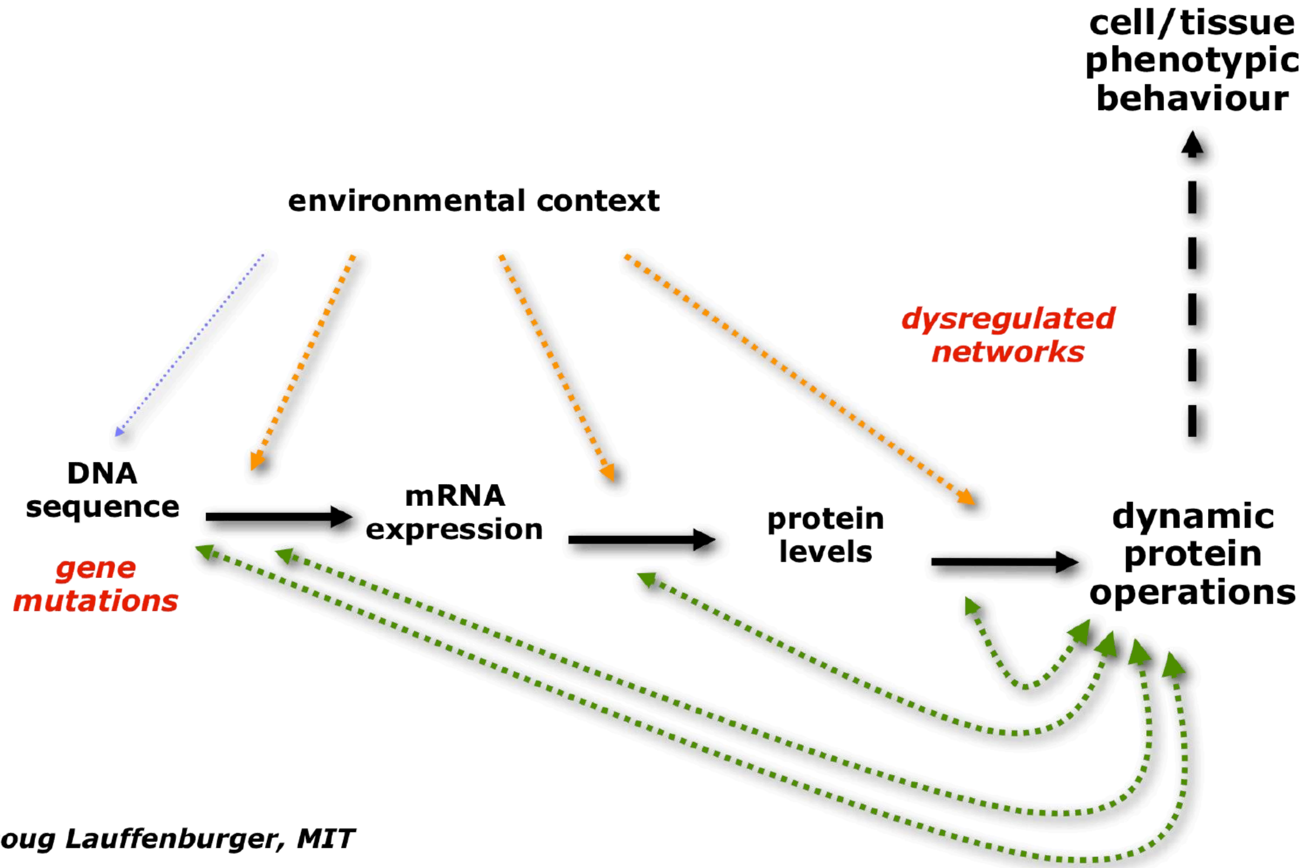
Get your T-Shirt!





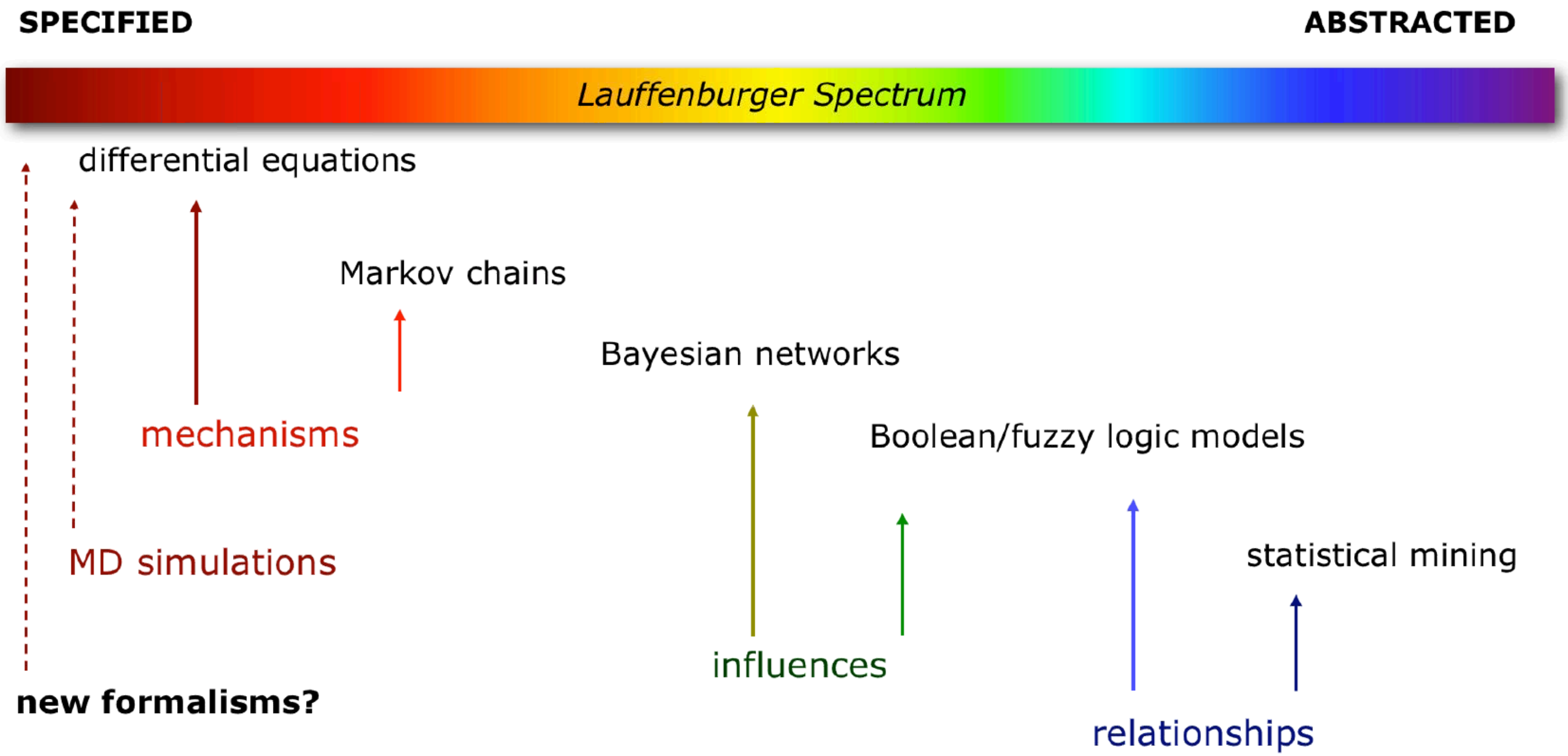
- Pretty simple, huh?
- NO! In reality, we know it’s very complex and we don’t know how to draw an accurate diagram.
- But our knowledge of biology at the molecular level comes from measuring these domains and formulating models of their interaction.
- In the last few years our ability to measure has exploded, producing awesome amounts of data.
- Analyzing that data has become a high performance computing problem, perhaps one of the largest and most important computing tasks facing us!

Our fundamental presumption is that disease has a molecular basis



Doug Lauffenburger, MIT

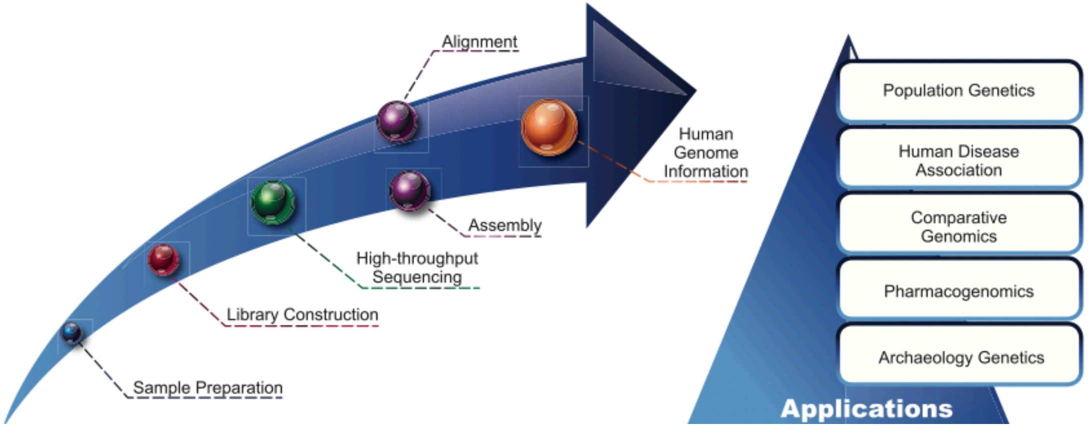
Computing in Biology has a very broad spectrum – we concentrate toward the right



Appropriate approach depends on question and data.
Computational Dynamic Range is Crucial.

Doug Lauffenburger, MIT.

Some workflows:
Human Population Genetics



Alignment

Human Genome Information

Assembly

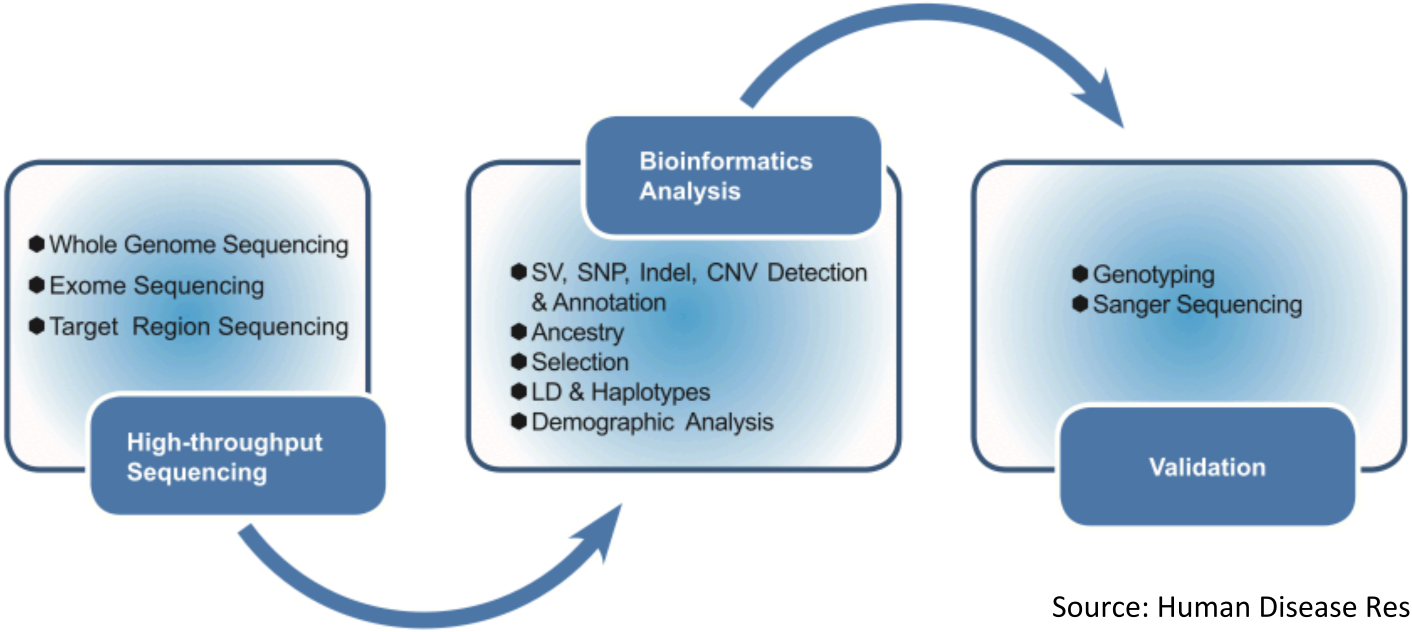
High-throughput Sequencing

Library Construction

Sample Preparation

- Population Genetics
- Human Disease Association
- Comparative Genomics
- Pharmacogenomics
- Archaeology Genetics

Applications



High-throughput Sequencing

- Whole Genome Sequencing
- Exome Sequencing
- Target Region Sequencing

Bioinformatics Analysis

- SV, SNP, Indel, CNV Detection & Annotation
- Ancestry
- Selection
- LD & Haplotypes
- Demographic Analysis

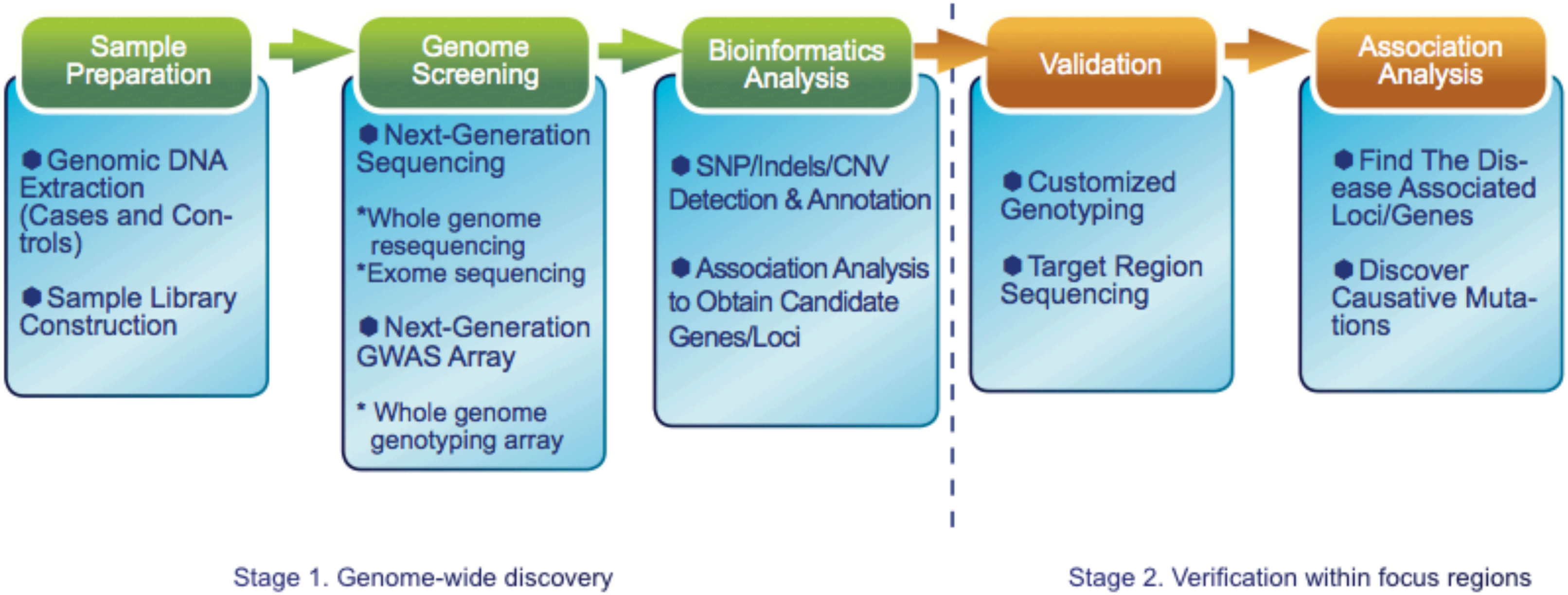
Validation

- Genotyping
- Sanger Sequencing

Source: Human Disease Research in the Era of Next-Generation Sequencing, BGI

Some workflows: Complex Disease Genome Wide Association Studies

- GWAS combines high-throughput sequencing and genotyping to uncover causative genetic mutations of complex human disease
- BGI's two stage workflow:



Sample Preparation
- Genomic DNA Extraction (Cases and Controls)
- Sample Library Construction

Genome Screening
- Next-Generation Sequencing
  - *Whole genome resequencing
  - *Exome sequencing
- Next-Generation GWAS Array
  - * Whole genome genotyping array

Bioinformatics Analysis
- SNP/Indels/CNV Detection & Annotation
- Association Analysis to Obtain Candidate Genes/Loci

Validation
- Customized Genotyping
- Target Region Sequencing

Association Analysis
- Find The Disease Associated Loci/Genes
- Discover Causative Mutations

Stage 1. Genome-wide discovery

Stage 2. Verification within focus regions

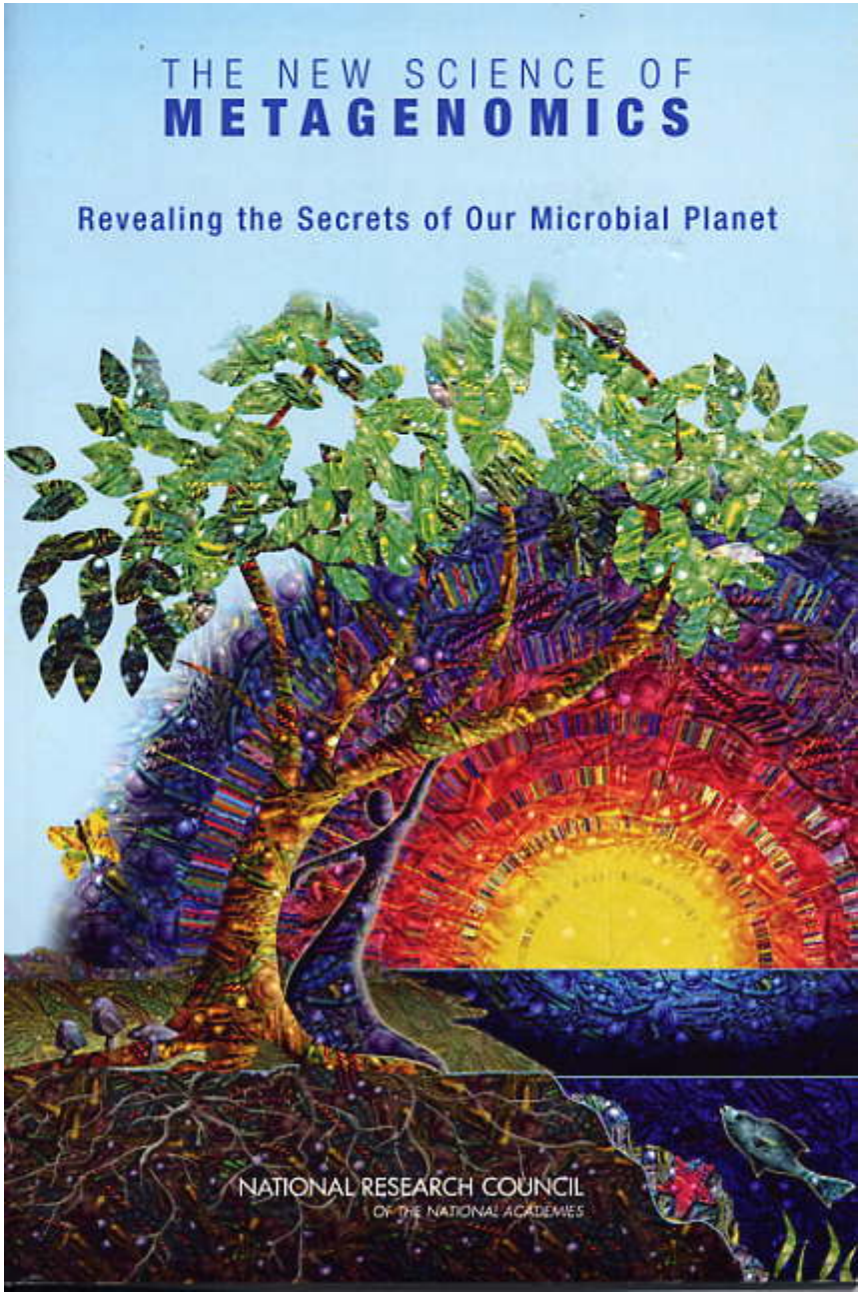
Figure 4. Workflow of the two-stage strategy

Source: Human Disease Research in the Era of Next-Generation Sequencing, BGI

Dealing with multiple organisms simultaneously

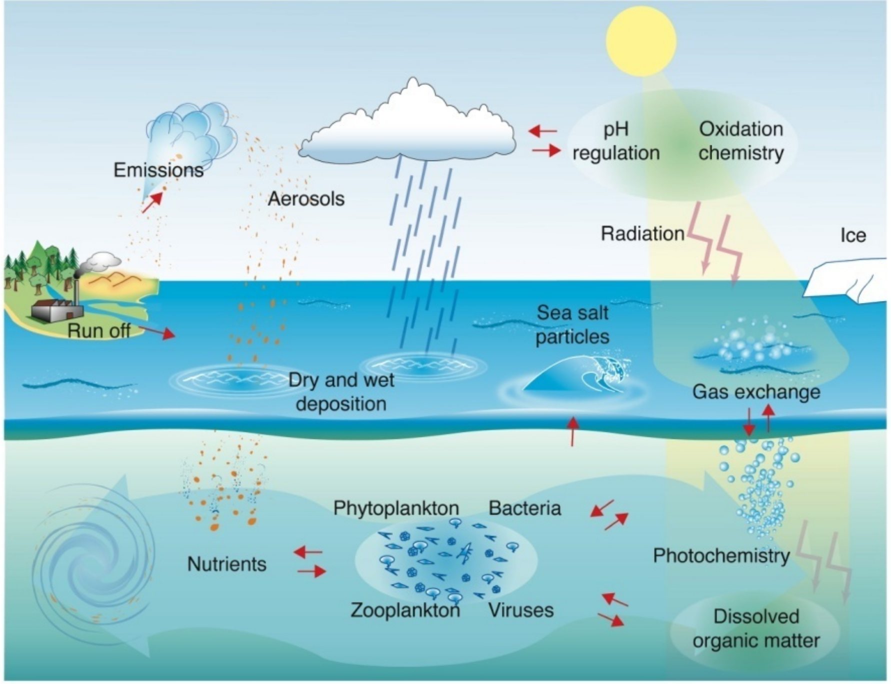
"Microbes run the world. It's that simple" American National Academies 2007

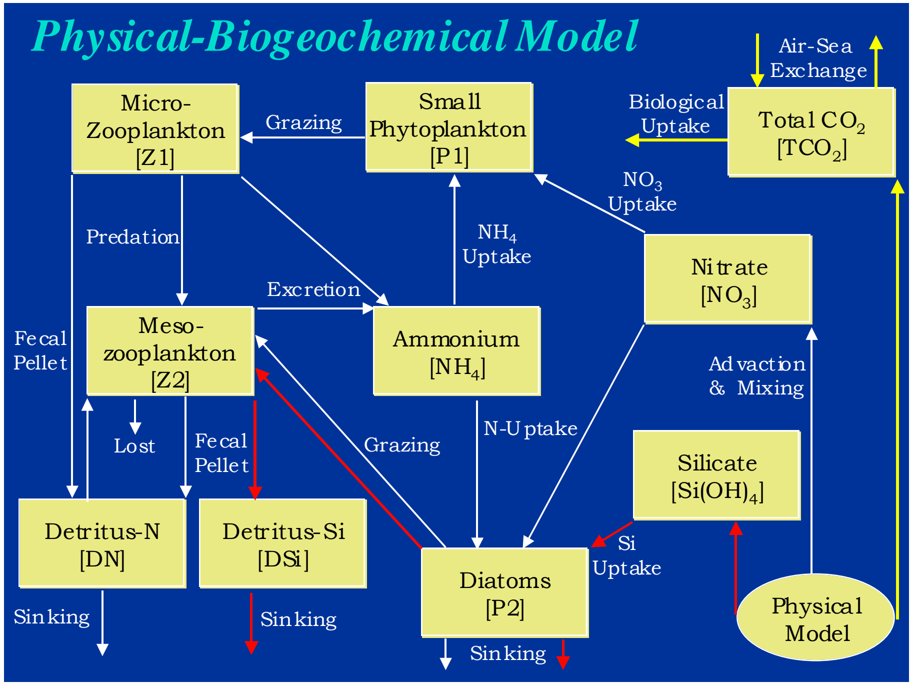
- Most (>99%) microbes cannot be studied in the laboratory
- *Understanding microbial communities in situ*
- *Who are they? and what do they do?*


THE NEW SCIENCE OF
METAGENOMICS
Revealing the Secrets of Our Microbial Planet
NATIONAL RESEARCH COUNCIL
OF THE NATIONAL ACADEMIES

http://www.nap.edu/catalog.php?record_id=11902

An example: one current “Biogeochemical Model” in global climate models





- Based upon: Moore, J. K., Doney, S. C., Kleypas, J. A., Glover, D. M., & Fung, I. Y.(2002) Deep-Sea Research Part II-Topical Studies in Oceanography 49, 403-46
- with added sulfur cycling and methane (for sea-bed methane release).

A very current example - the Human Microbiome Project

- Nature Volume 486 Number 7402 pp157-286 June 14, 2012 →
- The Human Microbiome Project Collection PLoS June 13, 2012
- NIH's site: http://hmpdacc.org/

What Analysis Tools are available to you?

Name	Nucleic Acid Population	Analysis Strategy
RNA-Seq	RNA (may be poly-A, mRNA, or total RNA)	Alignment of reads to “genes”; variations for detecting splice junctions and quantifying abundance
Small RNA Sequencing	Small RNA (often miRNA)	Alignment of reads to small RNA references (e.g. miRbase), then to the genome; quantify abundance
ChIP-Seq	DNA bound to protein, captured via antibody (ChIP=Chromatin ImmunoPrecipitation)	Align reads to reference genome, identify peaks and motifs
Structural Variation Analysis	Genomic DNA, with two reads (mate-pair reads) per DNA template	Align mate-pairs to reference sequence and interpret structural variants
De novo Sequencing	Genomic DNA, possibly with external data (e.g. cDNA, genomes of closely related species, etc.)	Piece together reads to assemble contigs, scaffolds, and (ideally) whole-genome sequence
Metagenomics	Entire RNA or DNA from a (usually microbial) community	Phylogenetic analysis of sequences

Genomics Sequencer that feed this analysis

- Illumina – Sequencing-by-synthesis chemistry; Generates 600 GB of data in a standard run lasting 27 hours
- Ion Torrent Systems (a subsidiary of Life Technologies) – Semiconductor based seq. First to announce $1000 genome
- Oxford Nanopore Technologies – Nanopore-based GridION technology; Enables direct sequencing of single DNA;
- Pacific Biosciences of California – Single molecule, real-time (SMRT) technology – real time analysis of biomolecules with single molecule resolution;
- 454 Life Sciences (a subsidiary of Roche Applied Sciences) – Target NGS of smaller exomes and genomes

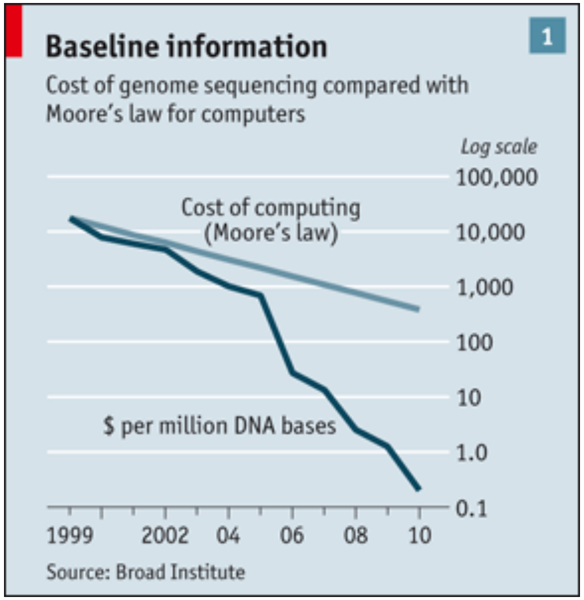
Genomic Sequencing: Market (source: IDC)

- Next Generation Sequencing (NGS) has become pervasive
  - Found in academic & clinical healthcare research
  - Agriculture & Pharmacology CROs quickly adopting
- Two primary business strategies for sequencing companies
  - Create & distribute sequencing instruments, and consumables
    - Highly specialized instruments require special training & care
    - Margin found in selling of consumables (e.g. printer & ink)
  - Sequencing Service Provider (SSP)
    - Biological sample sent to lab / Data is shipped to customer (often w/USB HD)
    - Lab provides different responses (raw data → fully analyzed)
- Illumina is current leader in sequencing instruments & technology
  - New technologies & companies coming on strong
    - Oxford Nanopore, Ion Torrent (owned by Life Tech), Pac-Bio …
- SSP is gaining traction leveraging cloud scale & capabilities
  - BGI, Complete Genomics (leveraging Illumina instruments) current leaders
  - Highly competitive space (many offerings all commoditizing price)

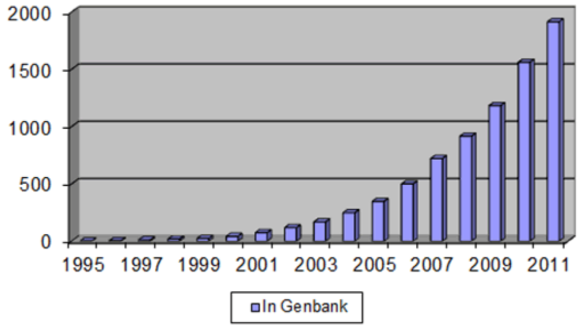
Genomic Sequencing: Market (source: IDC)

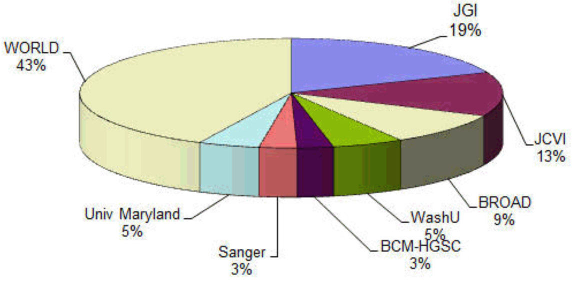
- Number of Gbase processed expected to grow dramatically
  - Sequencers produce 600+ Gbase/week
  - Expected to double in FY12
  - Raw & Analyzed storage ~250TB/Sequencer/YR
    - Anecdotal: “One customer is predicting increase to about 1.4M annual samples in 2014 from <100K today.”
  - Storage/Gbase decreasing due to new formats & improved compression
  - Cost of Gbase is shrinking - Human genome cost $50K in 2009 → $2.5K in 2012
- Continued growth in storage expected in sequencing workflows
  - High-end sequencer drives ~23.5TB of raw data/year
    - Analysis & archive drives ~10X that (~250TB/year)
- Sequencing requires compute & storage
    - Typical facility requires 250TB/sequencer/year (expected to grow next year)

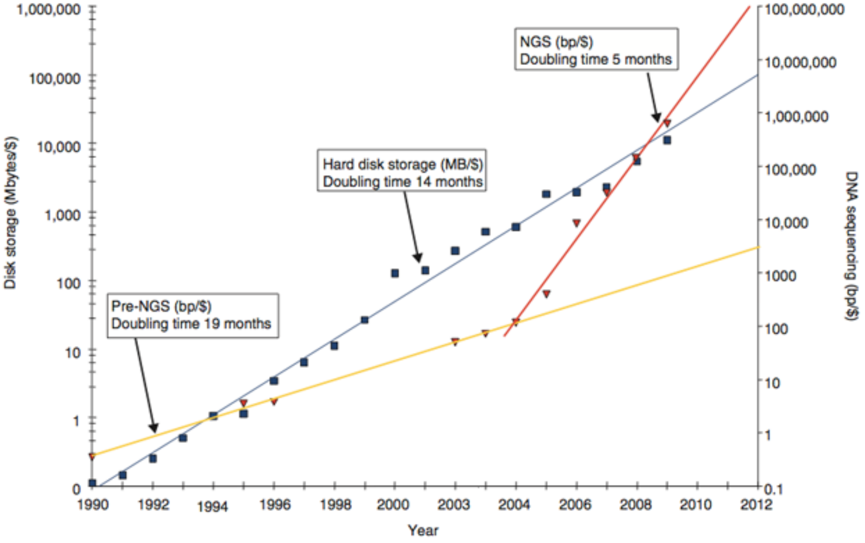
Some numbers illustrating the size of the problem



Complete genomes sequenced (Genomesonline.org – 7/5/12)





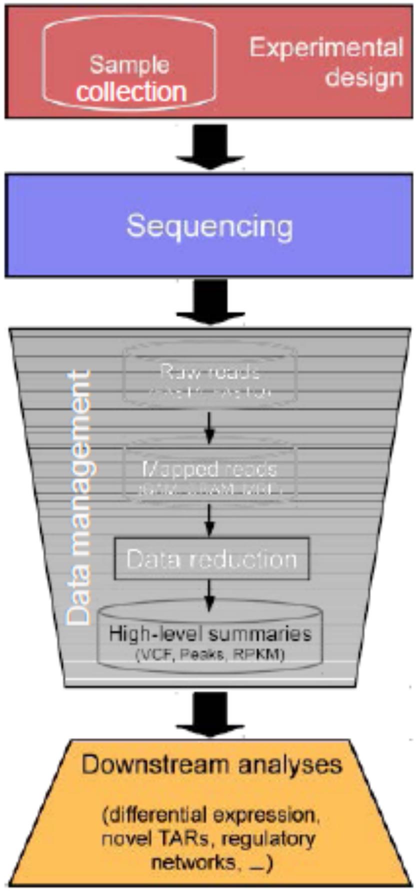


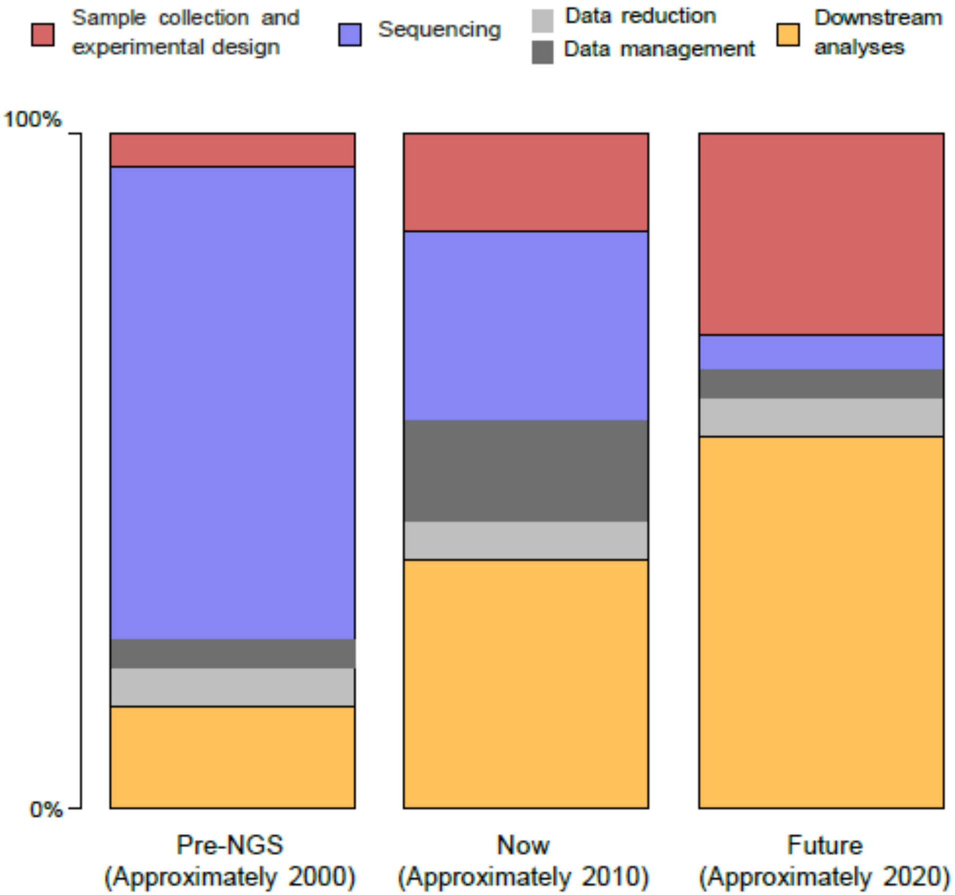
The impending collapse of the genome informatics ecosystem, Stein *Genome Biology* 2010 **11**:207

And … this data is geographically distributed

BGI demonstrated genomic data transfer at nearly 10 gigabits per second between US and China (http://phys.org/news/2012-06-bgi-genomic-gigabits-china.html)

Implications of Sequencing Productivity to Overall Costs





Total cost will be driven by Experimental Design and Compute/ Storage requirements (more than by sequencing itself)

Sboner, A, et al. 2011, ‘The real cost of sequencing: higher than you think!’, Genome Biology, 12:125

This data in perspective

- Two other well known “big data” fields
  - Astronomy/Astrophysics
    - Next generation space telescope
    - Square Kilometer Array
  - High energy physics
    - ATLAS
- This is Cool Stuff
  - But the size of data in these fields of study are dwarfed by the prospective magnitude of life sciences bioinformatics data … and processing requirements
  - Life Sciences directly touches our daily lives - $$

Break

After the break, let's look at some algorithms used to analyze this data

Many computing tasks in these workflows …

- Involve search and pattern matching
- There are many algorithms behind these tasks, but some important one appear often
  - Dynamic Programming ala Needleman- Wunsch and Smith-Waterman
  - Heuristic algorithms like BLAST and FASTA
  - The compression algorithm of Burrows-Wheeler
  - Directed graphs - deBruijn Graphs
- I will cover two general categories of applications that put severe demands on computing machinery
  - Sequence alignment
  - Genome Assembly

Alignment

- Pairwise alignment
  - Dynamic Programming methods
    - Global: Needleman-Wunsch
    - Local: Smith Waterman
  - Heuristic methods based on k-mers
    - BLAST
- Short Read Alignment
  - A compression algorithm
    - Burrows-Wheeler
- I’ll have to skip multiple sequence alignment for lack of time
  - Next year: BioHPC 102 ?

Sequence Alignment

- Detection of sequence similarity points to similarity in function or origin (homology) … or not!
- Comparison between two sequences (nucleic acid or protein) starts with a “pairwise alignment”
- Example:

```
             Sequence #1: CGGGTATCCAA
             Sequence #2: CCCTAGGTCCCA
```

```
Alignment #1 Sequence #1: CGGGTA--T-CCAA
             Sequence #2: CCC-TAGGTCCC-A
```

```
Alignment #2 Sequence #1: CGGGTA---TCCAA
             Sequence #2: CC--CTAGGTCCCA
```

```
Alignment #3 Sequence #1: C-GGGTA--TCCAA
             Sequence #2: CC--CTAGGTCCCA
```

Three alignments – which one is best?

Dynamic programming

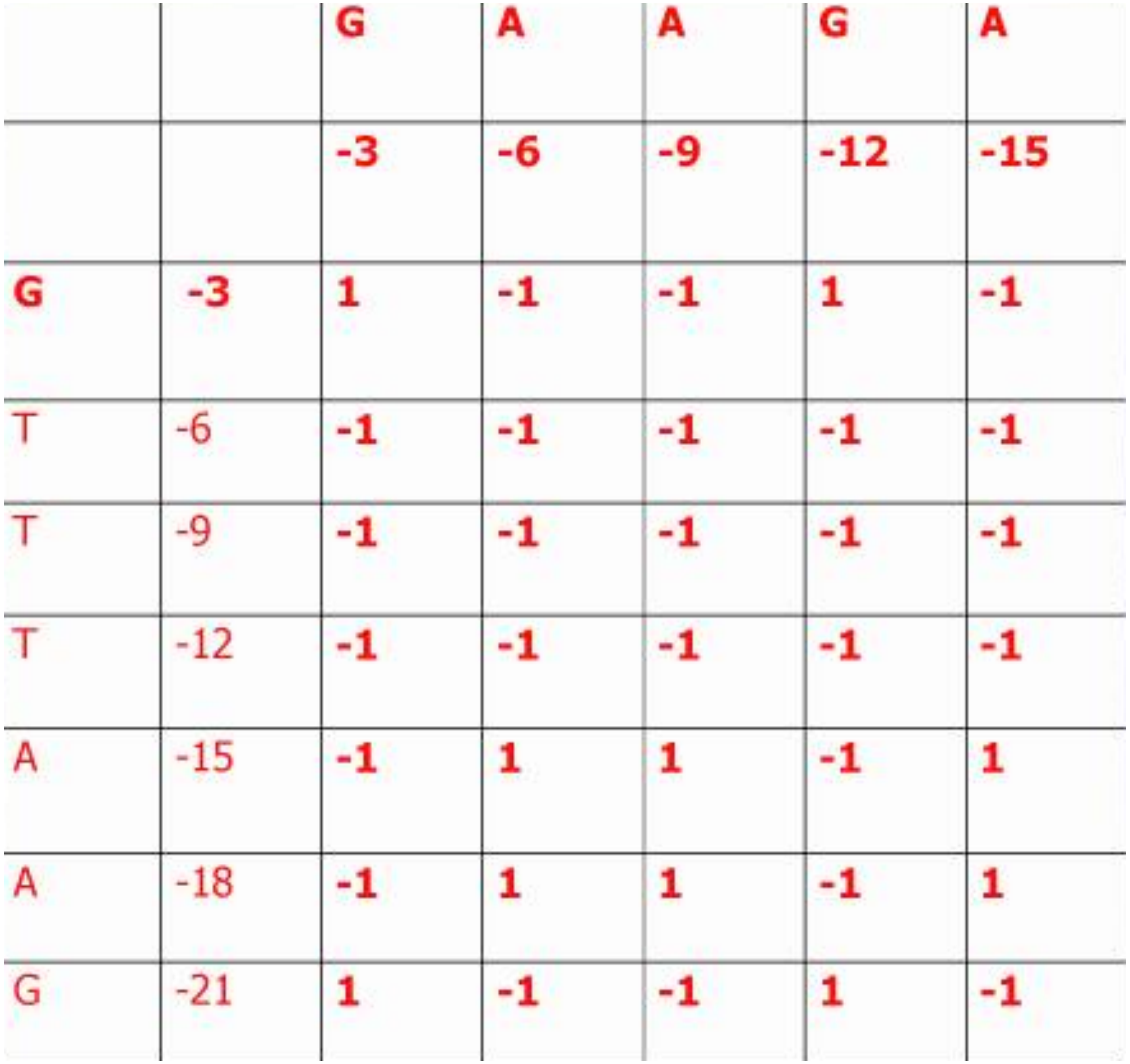
- Provides “optimal” alignment
- Recognizes (base or amino acid) matches, mismatches and gaps and “scores” each with a scoring rule or matrix
- Example: Find the global alignment of

  GAAGA and GTTTAAG

  – Use this rule for the initial alignment
    - Match = +1
    - Mismatch = -1
    - Gap = -3

  – Needleman-Wunsch – global alignment

Set up a matrix

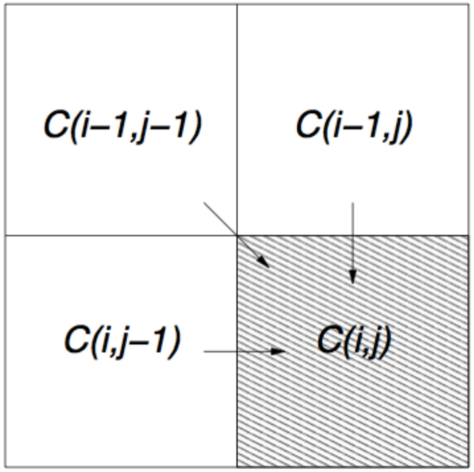


		G	A	A	G	A
		-3	-6	-9	-12	-15
G	-3	1	-1	-1	1	-1
T	-6	-1	-1	-1	-1	-1
T	-9	-1	-1	-1	-1	-1
T	-12	-1	-1	-1	-1	-1
A	-15	-1	1	1	-1	1
A	-18	-1	1	1	-1	1
G	-21	1	-1	-1	1	-1

Source: Incogen

NW Algorithm

- Use these rules
  - Move horizontally introducing a gap
    - Score += gap score
  - Move vertically introducing a gap
    - Score += gap score
  - Move diagonally
    - Score += corner value
- Then build a matrix of scores starting at the upper left
- Cell score is the maximum of q_{diag}, q_{up} or q_{left}
- Mark the path with an arrow back to the max-cell (if there is more than one, mark both)
- Then traceback through the arrows along the path is the alignment

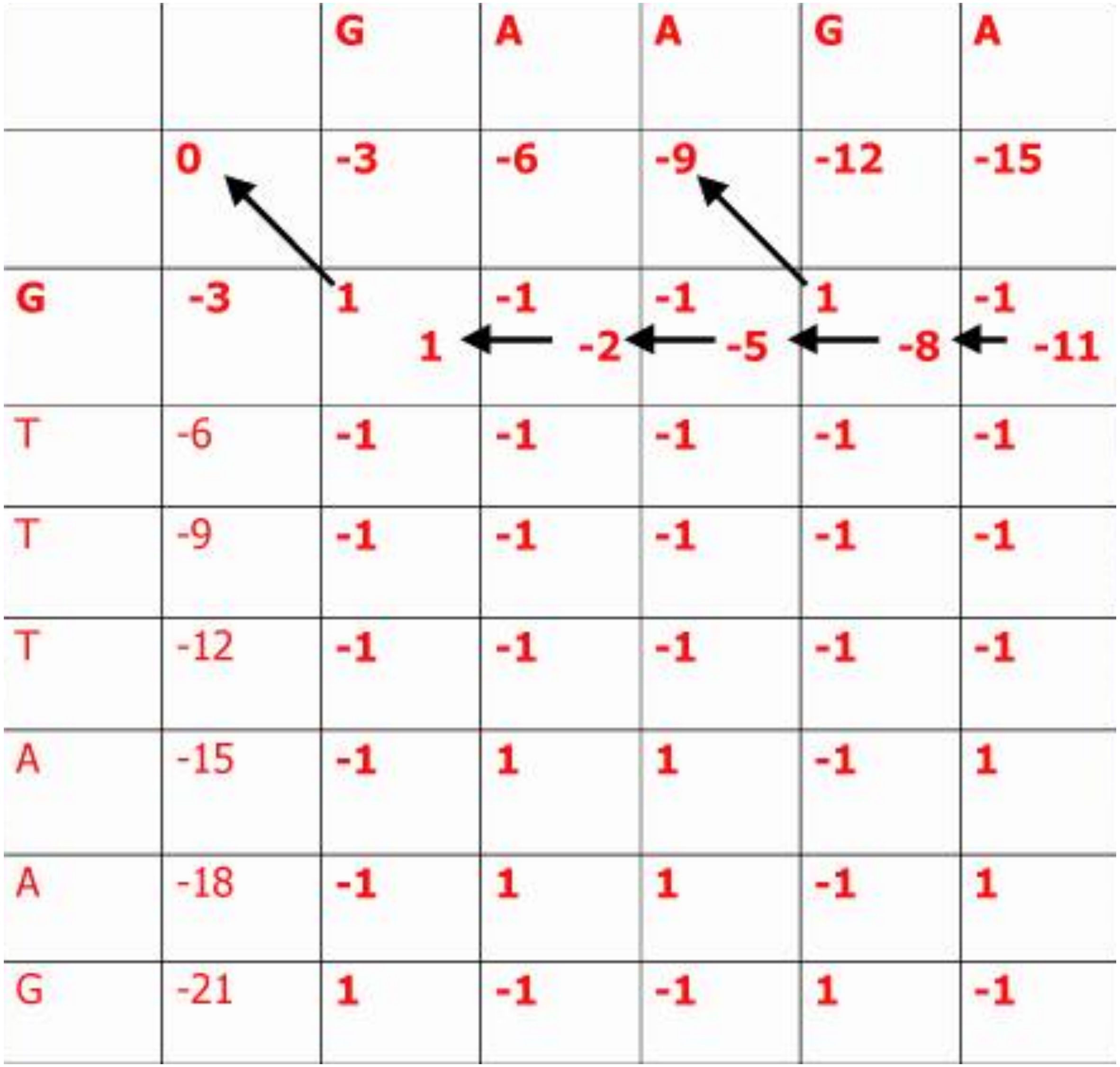


$$
\begin{aligned}
q_{diag} &= C(i-1,j-1) + S(i,j) \\
q_{up} &= C(i-1,j) + g \\
q_{left} &= C(i,j-1) + g
\end{aligned}
$$

Where

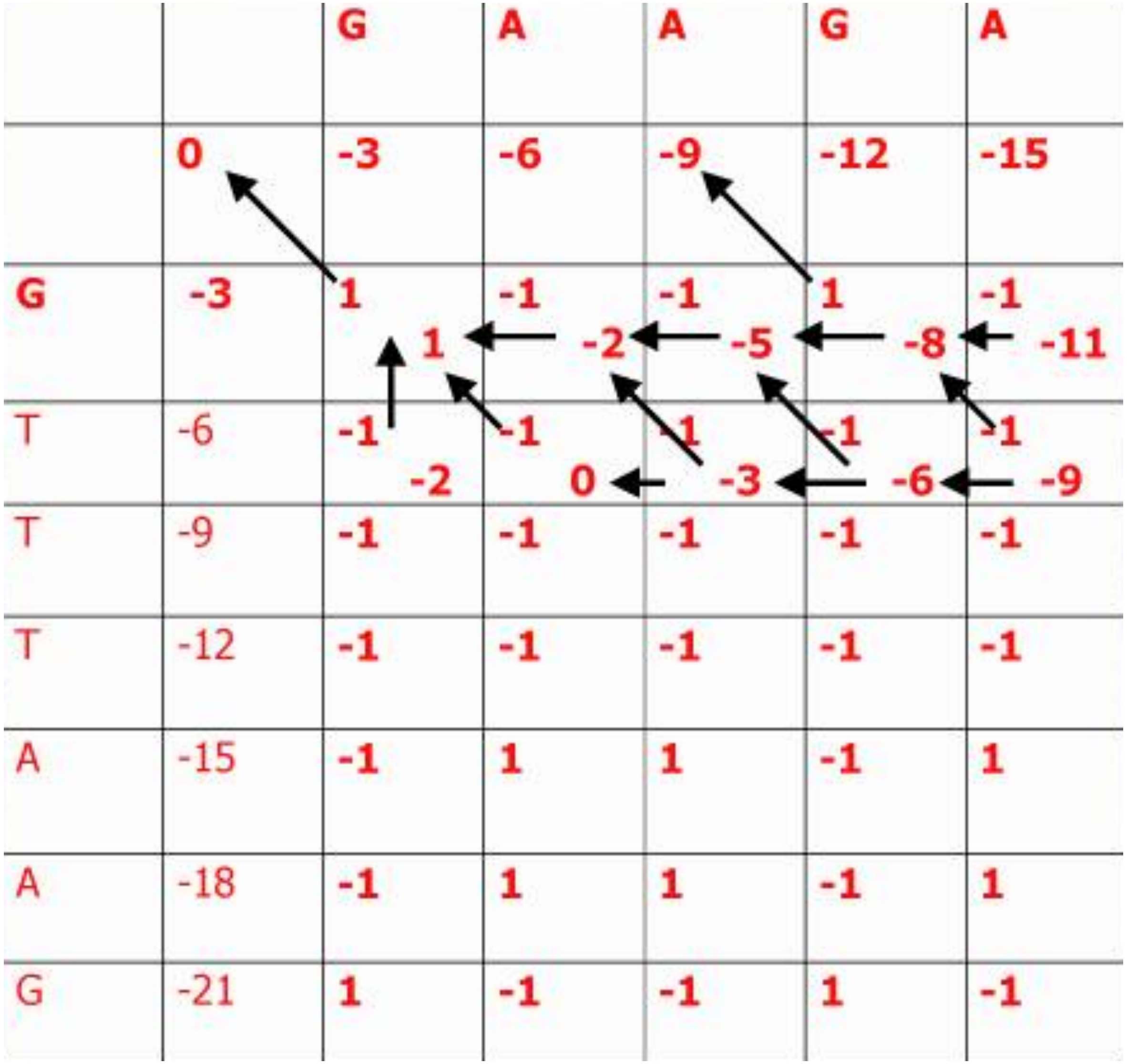
- C is the score previously calculated
- S is "substitution" score – in our case 1 or -1
- g is the gap score – in our case -3

Adapted from: Vladimir Likic Uni Melbourne



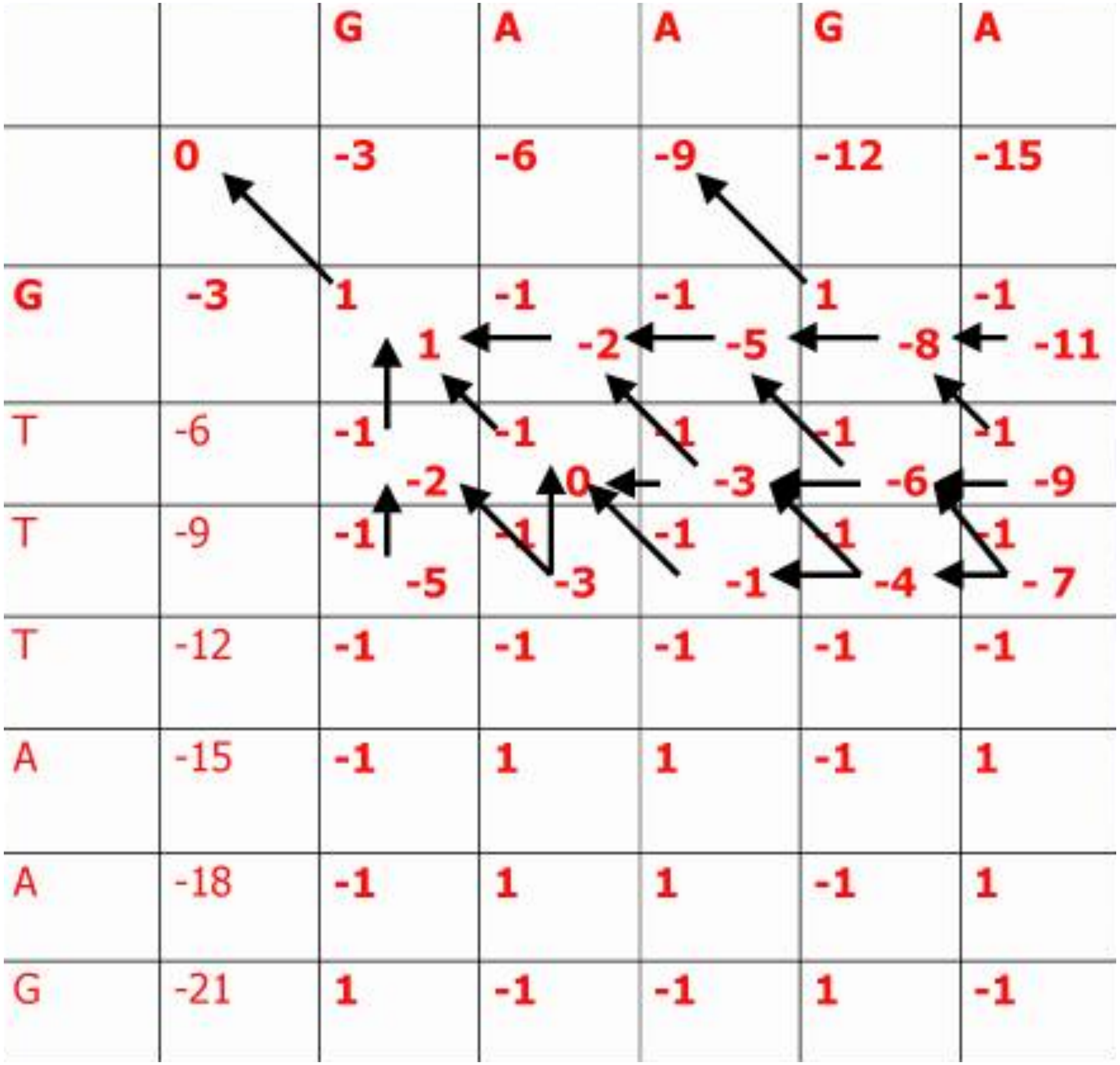
		G	A	A	G	A
	0	-3	-6	-9	-12	-15
G	-3	1 1	-1 -2	-1 -5	1 -8	-1 -11
T	-6	-1	-1	-1	-1	-1
T	-9	-1	-1	-1	-1	-1
T	-12	-1	-1	-1	-1	-1
A	-15	-1	1	1	-1	1
A	-18	-1	1	1	-1	1
G	-21	1	-1	-1	1	-1

Source: Incogen



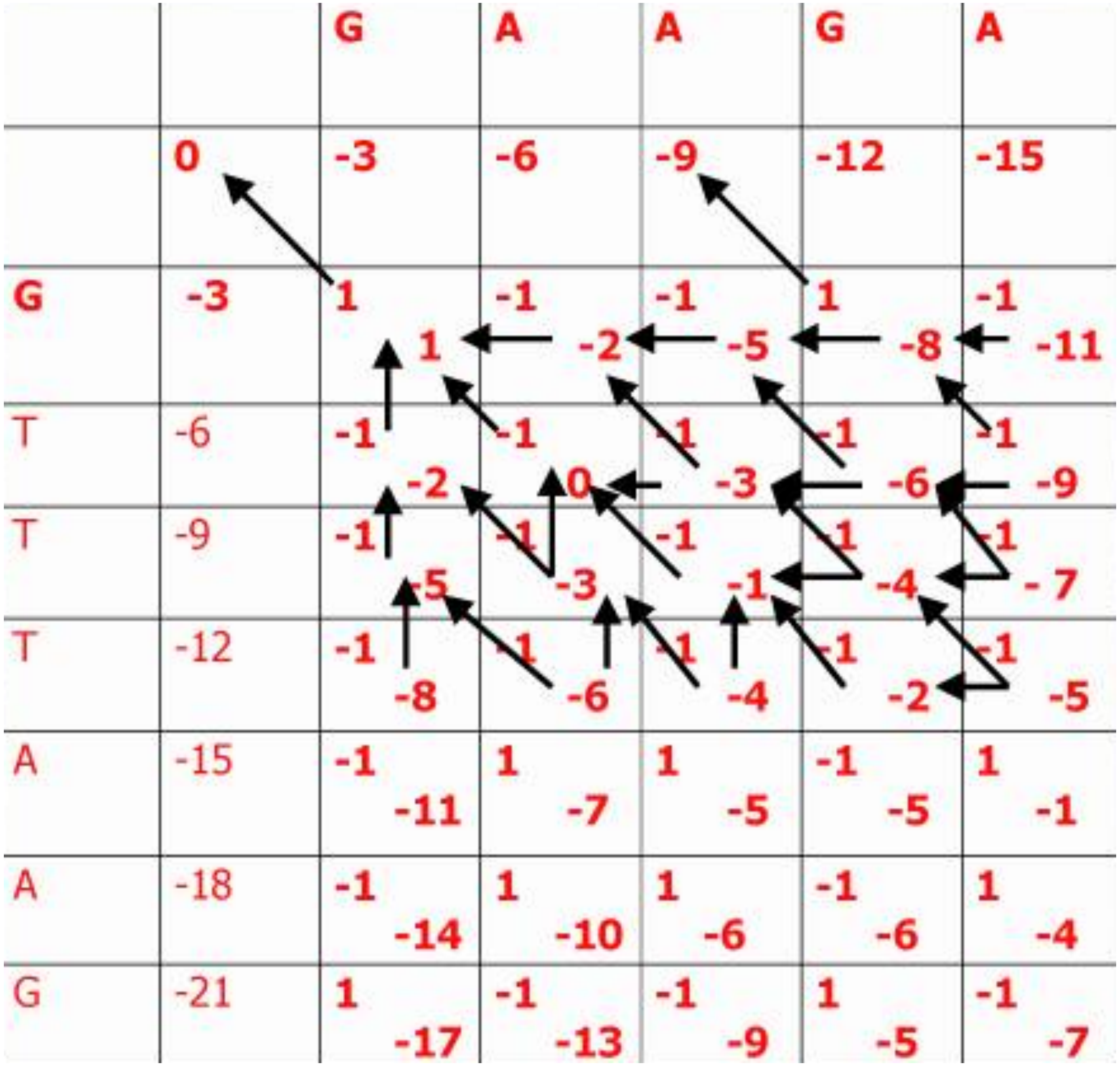
		G	A	A	G	A
	0	-3	-6	-9	-12	-15
G	-3	1 1	-1 -2	-1 -5	1 -8	-1 -11
T	-6	-1 -2	-1 0	-1 -3	-1 -6	-1 -9
T	-9	-1	-1	-1	-1	-1
T	-12	-1	-1	-1	-1	-1
A	-15	-1	1	1	-1	1
A	-18	-1	1	1	-1	1
G	-21	1	-1	-1	1	-1

Source: Incogen



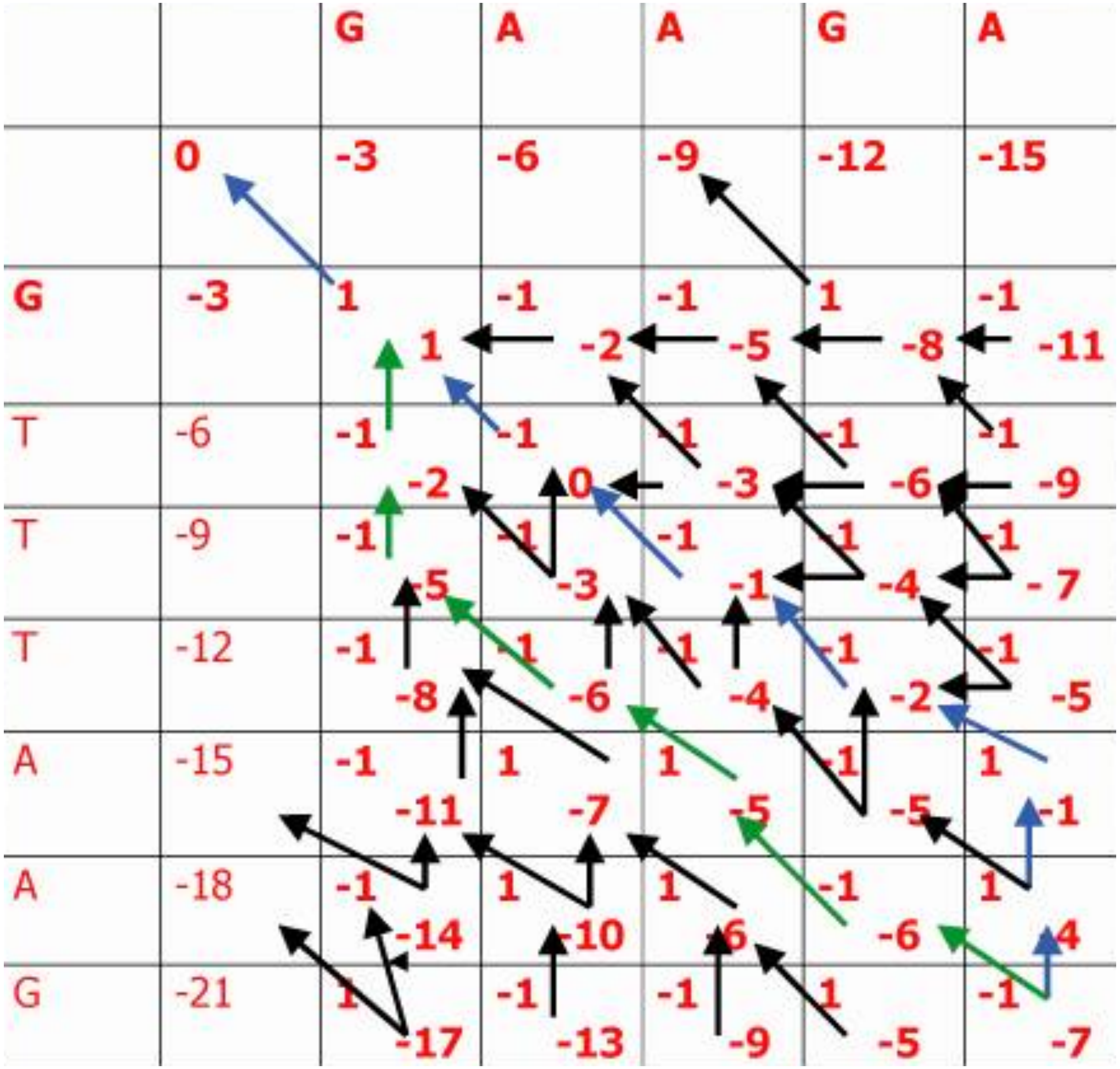
		G	A	A	G	A
	0	-3	-6	-9	-12	-15
G	-3	1	-1	-1	1	-1
T	-6	-1	-1	-1	-1	-1
T	-9	-1	-1	-1	-1	-1
T	-12	-1	-1	-1	-1	-1
A	-15	-1	1	1	-1	1
A	-18	-1	1	1	-1	1
G	-21	1	-1	-1	1	-1

Source: Incogen



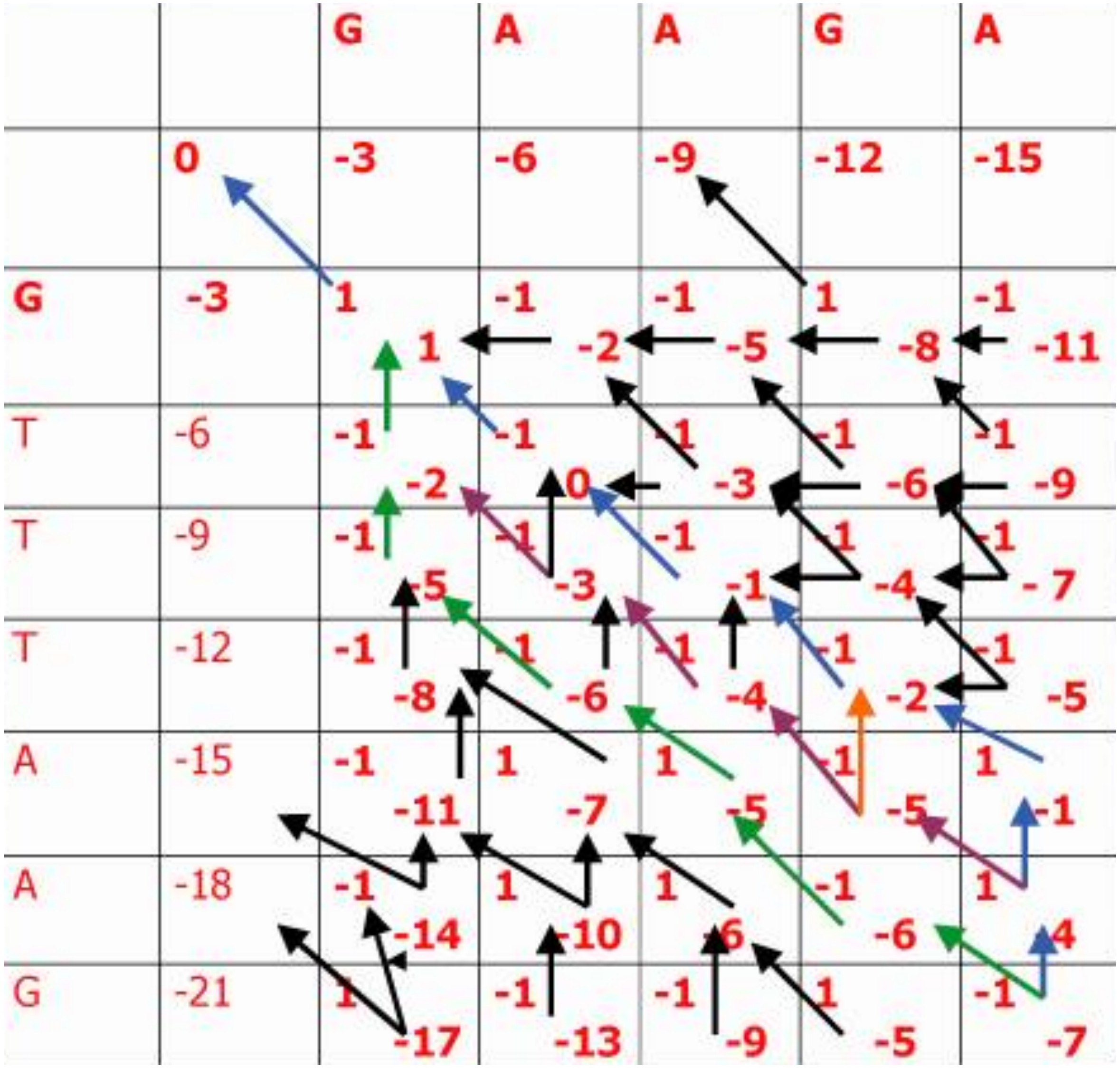
		G	A	A	G	A
	0	-3	-6	-9	-12	-15
G	-3	1 1	-1 -2	-1 -5	1 -8	-1 -11
T	-6	-1 -2	-1 0	-1 -3	-1 -6	-1 -9
T	-9	-1 -5	-1 -3	-1 -1	-1 -4	-1 - 7
T	-12	-1 -8	-1 -6	-1 -4	-1 -2	-1 -5
A	-15	-1 -11	1 -7	1 -5	-1 -5	1 -1
A	-18	-1 -14	1 -10	1 -6	-1 -6	1 -4
G	-21	1 -17	-1 -13	-1 -9	1 -5	-1 -7

Source: Incogen



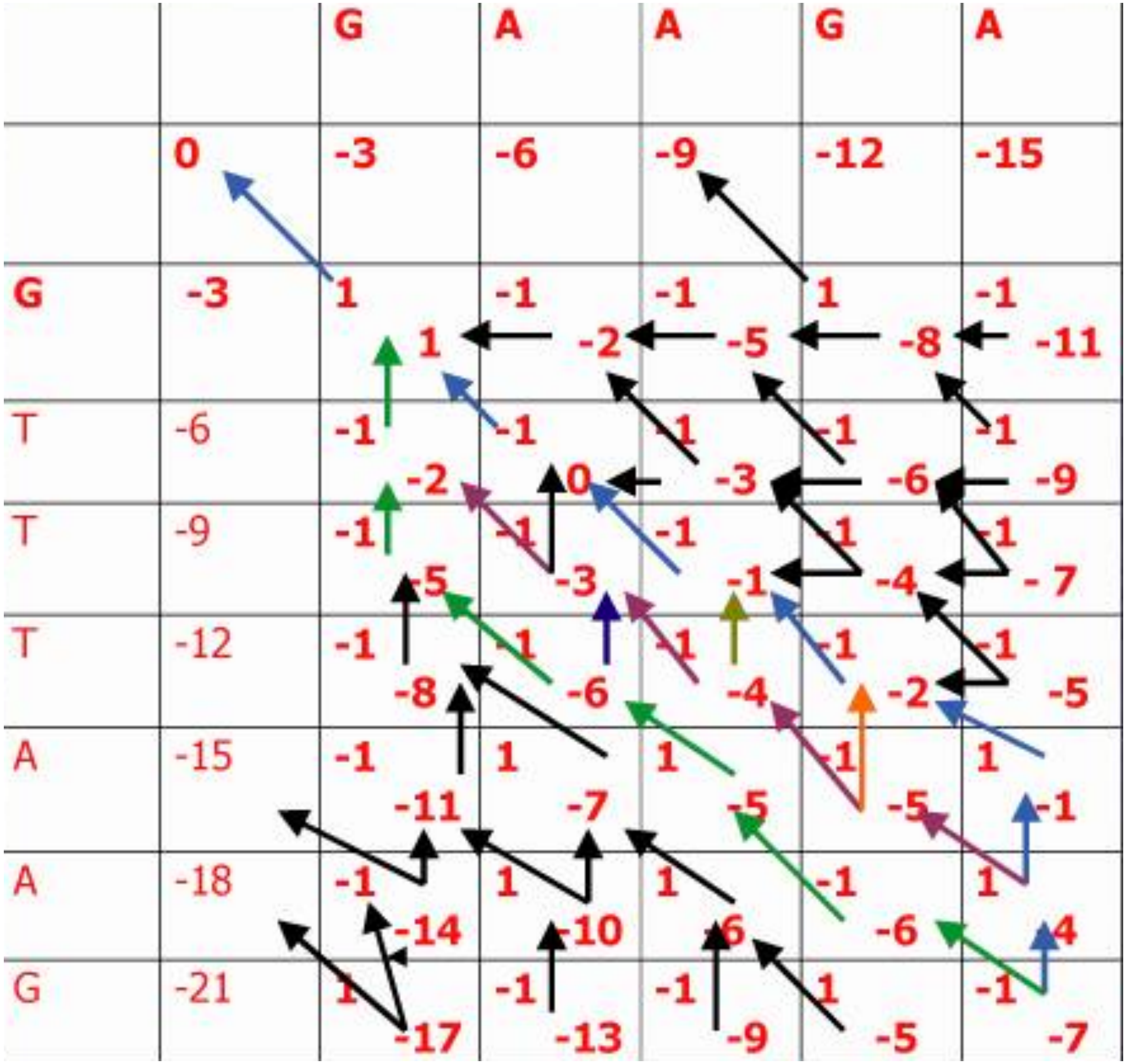
		G	A	A	G	A
	0	-3	-6	-9	-12	-15
G	-3	1 1	-1 -2	-1 -5	1 -8	-1 -11
T	-6	-1 -2	-1 0	-1 -3	-1 -6	-1 -9
T	-9	-1 -5	-1 -3	-1 -1	-1 -4	-1 - 7
T	-12	-1 -8	-1 -6	-1 -4	-1 -2	-1 -5
A	-15	-1 -11	1 -7	1 -5	-1 -5	1 -1
A	-18	-1 -14	1 -10	1 -6	-1 -6	1 -4
G	-21	1 -17	-1 -13	-1 -9	1 -5	-1 -7

Source: Incogen



		G	A	A	G	A
	0	-3	-6	-9	-12	-15
G	-3	1 1	-1 -2	-1 -5	1 -8	-1 -11
T	-6	-1 -2	-1 0	-1 -3	-1 -6	-1 -9
T	-9	-1 -5	-1 -3	-1 -1	-1 -4	-1 - 7
T	-12	-1 -8	-1 -6	-1 -4	-1 -2	-1 -5
A	-15	-1 -11	1 -7	1 -5	-1 -5	1 -1
A	-18	-1 -14	1 -10	1 -6	-1 -6	1 -4
G	-21	1 -17	-1 -13	-1 -9	1 -5	-1 -7

Source: Incogen



		G	A	A	G	A
	0	-3	-6	-9	-12	-15
G	-3	1 1	-1 -2	-1 -5	1 -8	-1 -11
T	-6	-1 -2	-1 0	-1 -3	-1 -6	-1 -9
T	-9	-1 -5	-1 -3	-1 -1	-1 -4	-1 - 7
T	-12	-1 -8	-1 -6	-1 -4	-1 -2	-1 -5
A	-15	-1 -11	1 -7	1 -5	-1 -5	1 -1
A	-18	-1 -14	1 -10	1 -6	-1 -6	1 -4
G	-21	1 -17	-1 -13	-1 -9	1 -5	-1 -7

Source: Incogen

Global alignments with the highest score

```
GAAGA--
           = -7
GTTTAAG
```

```
G-A-AGA
           = -7
GTTTAAG
```

```
G--AAGA
           = -7
GTTTAAG
```

```
G-AAGA-
           = -7
GTTTAAG
```

```
GAAA-G-
           = -7
GTTTAAG
```

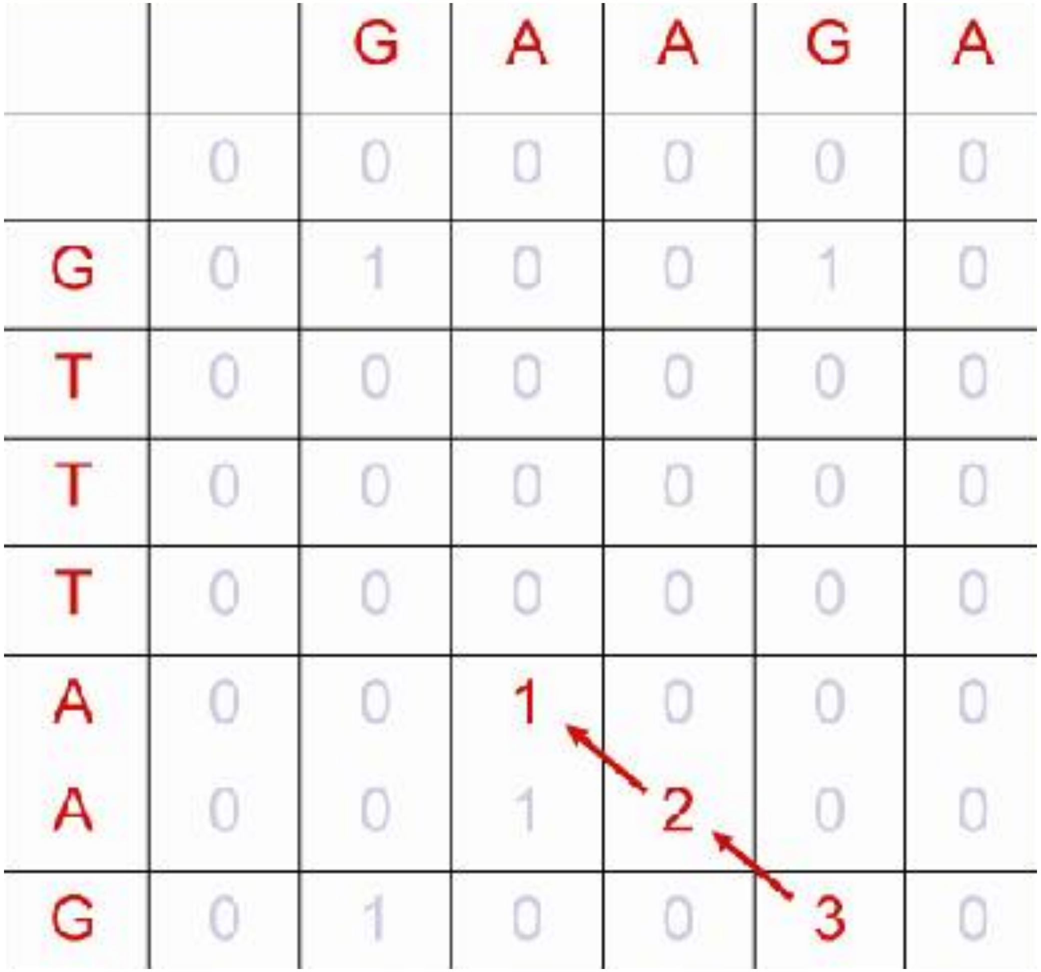
```
GAA-GA-
           = -7
GTTTAAG
```

- There are six global alignments with score = -7!
- But it might make sense to align only certain regions, perhaps those that are conserved → local alignment

Smith Waterman – Local alignment

- A simple variation of Needleman-Wunsch
  - local alignment: mismatches and gaps at the beginning and end of the sequences score 0.
  - Replaces the overall score by zero if it becomes negative values for all alternative pathways.
  - Find the highest score in the table and trace the path back until we come to a cell with a score of zero - this cell is not included in the alignment
- This simple approach restricts alignment to regions of reasonably high similarity.
- However "exact" alignment algorithms are so computationally expensive that they become unrealistically slow if one wants to compare a sequence to a background database of sequences.
- Thus the much wider use of heuristic algorithms like FASTA and BLAST

```
   GAAGA
              =  3
GTTTAAG
```



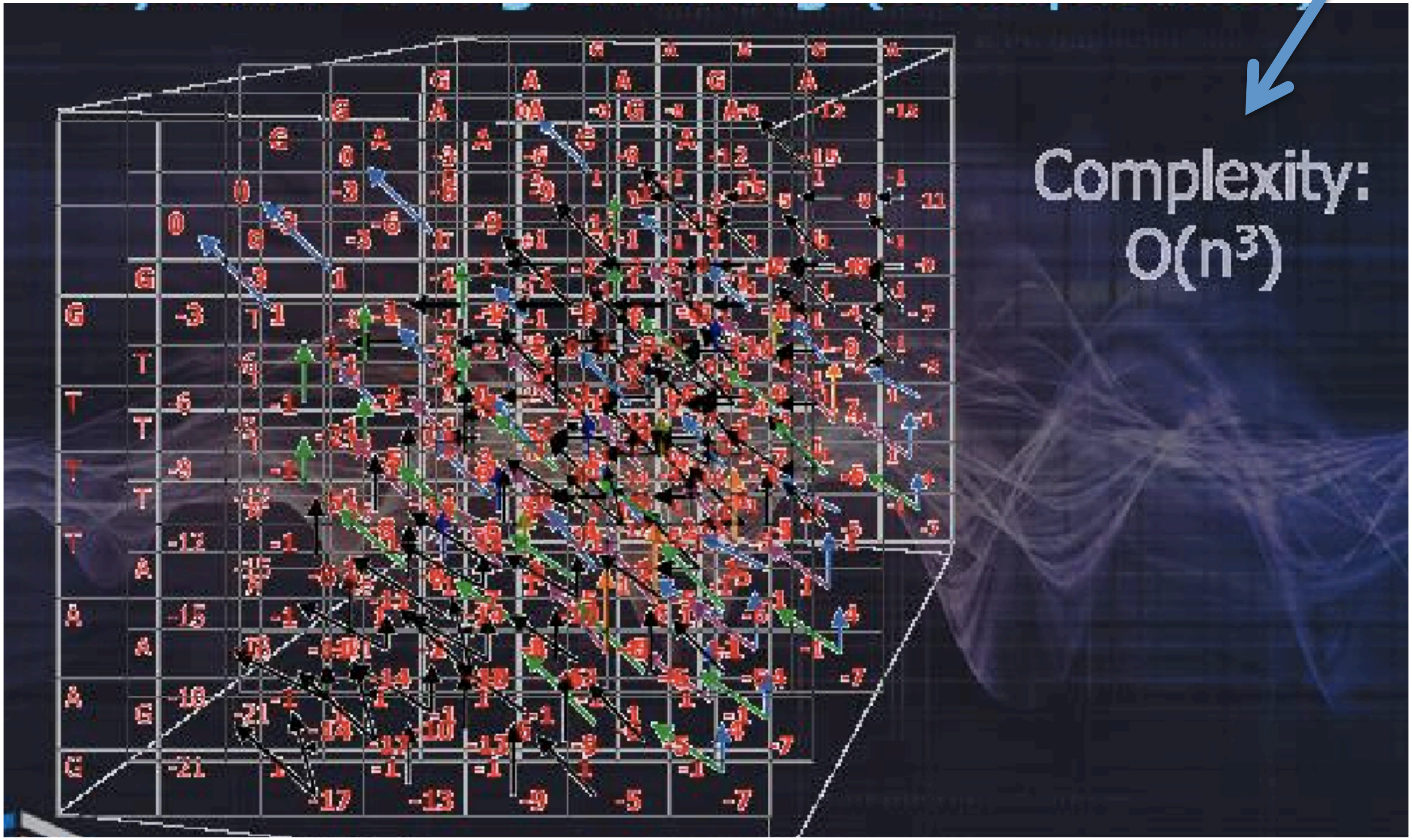
		G	A	A	G	A
	0	0	0	0	0	0
G	0	1	0	0	1	0
T	0	0	0	0	0	0
T	0	0	0	0	0	0
T	0	0	0	0	0	0
A	0	0	1	0	0	0
A	0	0	1	2	0	0
G	0	1	0	0	3	0

Adapted from "Sequence Alignment in HIV Computational Analysis"
Ana Abecasis, Anne-Mieke Vandamme, and Philippe Lemey and from Incogen

Dynamic Programming becomes quickly intractable for multiple alignment

Schematic of DP on multiple alignment of 3 sequences

Note that



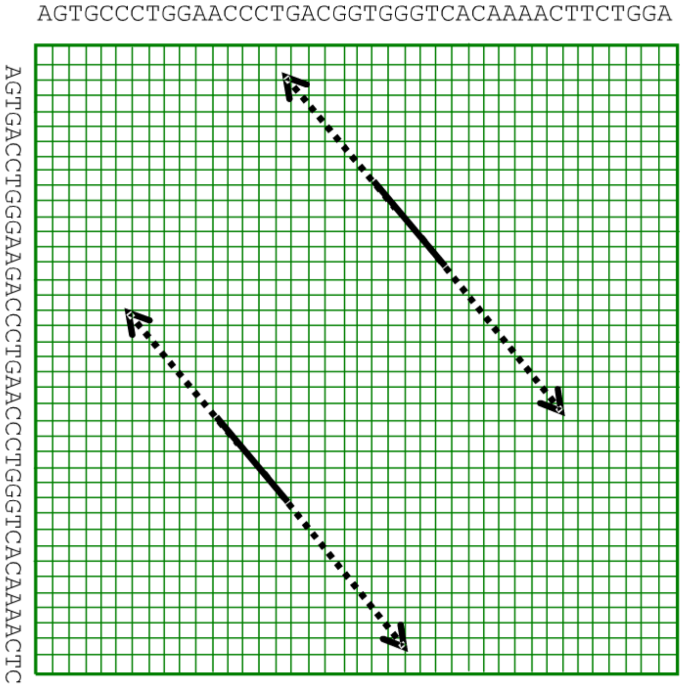
Source: Incogen

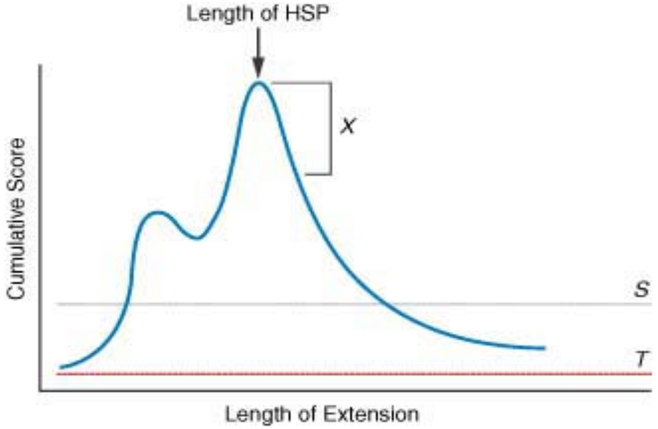
Heuristic Algorithms

- Tries out most likely alignments by
  - Finding all k-mers (words) of length l or larger in both sequences
  - When finding a perfect match the alignment is extended until:
    - Either sequence ends or
    - The score drops below a threshold
- BLAST (and FASTA) are much faster than NW or SW, but less sensitive
  - Let’s look at BLAST

BLAST – The most used app

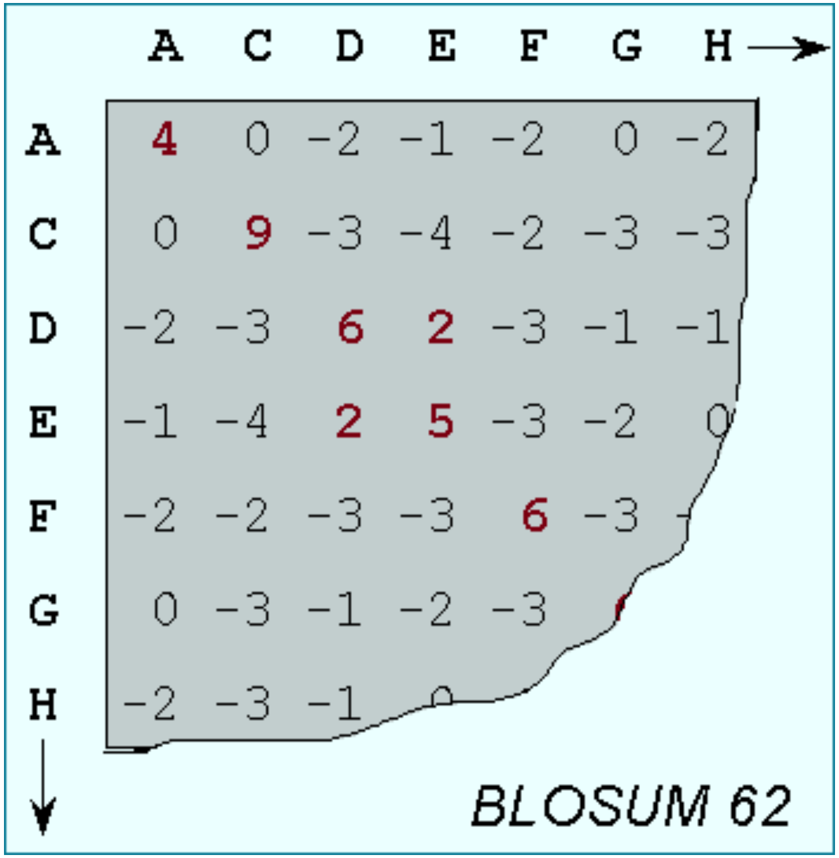
- Basic Local Alignment Search Tools are a set of sequence comparison algorithms introduced in 1990 that are used to search sequence databases for optimal local alignments to a query.
    - Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) "Basic local alignment search tool." J. Mol. Biol. 215:403-410.
    - Altschul SF, Madden TL, Schaeffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." NAR 25:3389-3402.
- Scoring of matches done using scoring matrices
- Sequences are split into words (default n=3)
- BLAST algorithm extends the initial "seed" hit into an "high scoring segment pair"
    - HSP = high scoring segment pair = Local optimal alignment
- Word hits are then extended in either direction in an attempt to generate an alignment with a score exceeding the threshold of "S".


AGTGCCCTGGAACCCTGACGGTGGGTCACAAAACTTCTGGA
AGTGACCCTGGGAAGACCCTGAACCCTGGGTCACAAAACTC



What’s a scoring matrix?

- Substitution matrices are used for amino acid alignments.
  - each possible residue substitution is given a score
- A simpler unitary matrix is used for DNA pairs (+1 for match, -2 mismatch)
- These substitutions are motived by a simple probabilistic model of the likelihood of one residue being replaced by another
  - Derived from experiment
  - There are numerous scoring matrices



	A	C	D	E	F	G	H
A	4	0	-2	-1	-2	0	-2
C	0	9	-3	-4	-2	-3	-3
D	-2	-3	6	2	-3	-1	-1
E	-1	-4	2	5	-3	-2	0
F	-2	-2	-3	-3	6	-3	-
G	0	-3	-1	-2	-3	6	
H	-2	-3	-1	0			

BLOSUM 62

Pick your BLAST
There's a bunch

- **Blastp:** Compares an amino acid query sequence against a protein sequence database.
- **Blastn:** Compares a nucleotide query sequence against a nucleotide sequence database.
- **Blastx**: Compares a nucleotide query sequence translated in all reading frames against a protein sequence database. You could use this option to find potential translation products of an unknown nucleotide sequence.
- **Tblastn**: Compares a protein query sequence against a nucleotide sequence database dynamically translated in all reading frames.
- **Tblastx**: Compares the six-frame translations of a nucleotide query sequence against the six-frame translations of a nucleotide sequence database.
- **Megablast**: Contiguous Nearly identical sequences and Discontiguous Cross-species comparison
- **PSI-BLAST**: Position Specific. Automatically generates a position specific score matrix (PSSM)
- **RPS-BLAST**: Position Specific. Searches a database of PSI-BLAST PSSMs
- … And more (MPI-BLAST, etc.)

Computing limitations

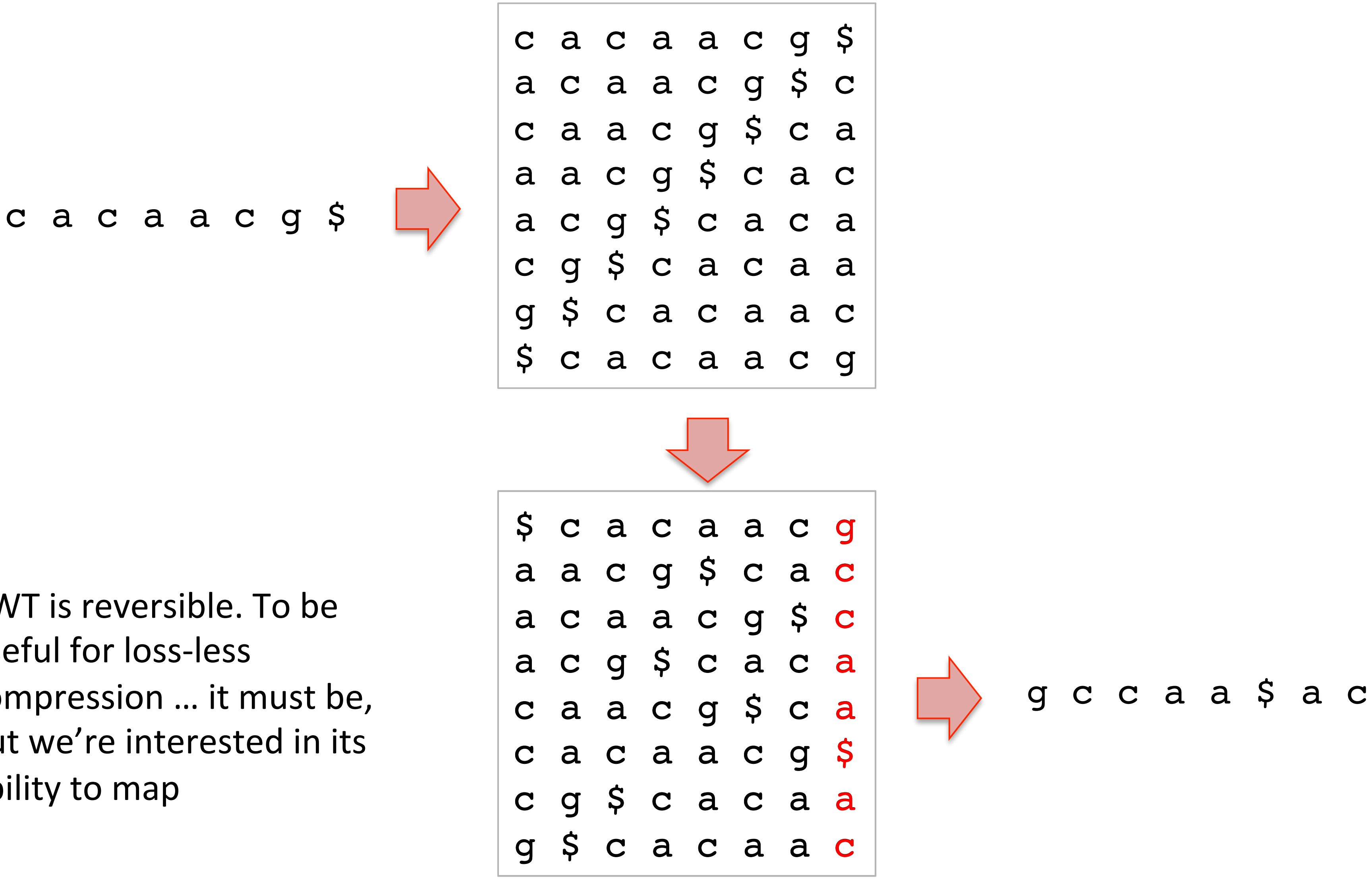
- Common characteristic: single threaded, single address space
- Generalize to a cluster, multiple threads and distributed address space. Examples:
  - MPI-Blast (www.mpiBlast.org), uses standard message passing to distribute compute threads.
    - Projects exist to implement on Amazon EC2
  - BlastReduce → Cloudburst (**CloudBurst: highly sensitive read mapping with MapReduce, M. Schatz,** Bioinformatics (2009) 25 (11):1363-1369) uses mapreduce and has been implemented in hadoop (actually Amazon's Map-Reduce) on Amazon EC2.
    - Watch for this connection later in the talk

pause

Short Read Aligners

- Why? → changes in technology
  - HT Sequencers produce vast numbers of reads
  - These reads are shorter than the previous Sanger technology
  - This combination renders older aligners (BLAST etc.) inefficient and
- A current list (as of June 18, 2012)
  - http://en.wikipedia.org/wiki/List_of_sequence_alignment_software#Short-Read_Sequence_Alignment
- The Burrows-Wheeler *text compression* algorithm proves useful
  - (Burrows M and Wheeler D (1994), *A block sorting lossless data compression algorithm*, Technical Report 124, Digital Equipment Corporation)
  - This is the underlying algorithm of bzip and bzip2.
  - Text compression? What does that have to do with alignment?
  - Combine B-W's reversible compression with the matching algorithm algorithm of Ferragina and Manzini (Opportunistic data structures with applications. Proceedings of the 41st Annual Symposium on Foundations of Computer Science. IEEE Computer Society; 2000.)

The Burrows-Wheeler transform (1994; 1983)



c a c a a c g $

c	a	c	a	a	c	g	$
a	c	a	a	c	g	$	c
c	a	a	c	g	$	c	a
a	a	c	g	$	c	a	c
a	c	g	$	c	a	c	a
c	g	$	c	a	c	a	a
g	$	c	a	c	a	a	c
$	c	a	c	a	a	c	g

$	c	a	c	a	a	c	g
a	a	c	g	$	c	a	c
a	c	a	a	c	g	$	c
a	c	g	$	c	a	c	a
c	a	a	c	g	$	c	a
c	a	c	a	a	c	g	$
c	g	$	c	a	c	a	a
g	$	c	a	c	a	a	c

g c c a a $ a c

BWT is reversible. To be useful for loss-less compression ... it must be, but we're interested in its ability to map

Source: Mott et. al., Wellcome Trust

The “Last-First mapping” property

The relative ordering of a particular character (say c) in column 1 is the same as that in the last column

c_1 a c_2 a a c_3 g $\$$

$\$$	c_1	a	c_2	a	a	c_3	g
a	a	c_3	g	$\$$	c_1	a	c_2
a	c_2	a	a	c_3	g	$\$$	c_1
a	c_3	g	$\$$	c_1	a	c_2	a
c_2	a	a	c_3	g	$\$$	c_1	a
c_1	a	c_2	a	a	c_3	g	$\$$
c_3	g	$\$$	c_1	a	c_2	a	a
g	$\$$	c_1	a	c_2	a	a	c_3

Source: Mott et. al., Wellcome Trust

The “Last-First mapping” property

The relative ordering of a particular character (say c) in column 1 is the same as that in the last column

$c_1\ a\ c_2\ a\ a\ c_3\ g\ \$$

$$\begin{array}{cccccccc}
\$ & c_1 & a & c_2 & a & a & c_3 & g \\
a & a & c_3 & g & \$ & c_1 & a & c_2 \\
a & c_2 & a & a & c_3 & g & \$ & c_1 \\
a & c_3 & g & \$ & c_1 & a & c_2 & a \\
c_2 & a & a & c_3 & g & \$ & c_1 & a \\
c_1 & a & c_2 & a & a & c_3 & g & \$ \\
c_3 & g & \$ & c_1 & a & c_2 & a & a \\
g & \$ & c_1 & a & c_2 & a & a & c_3
\end{array}$$

Source: Mott et. al., Wellcome Trust

After you have indexed the first string

- Let’s look up and match a query

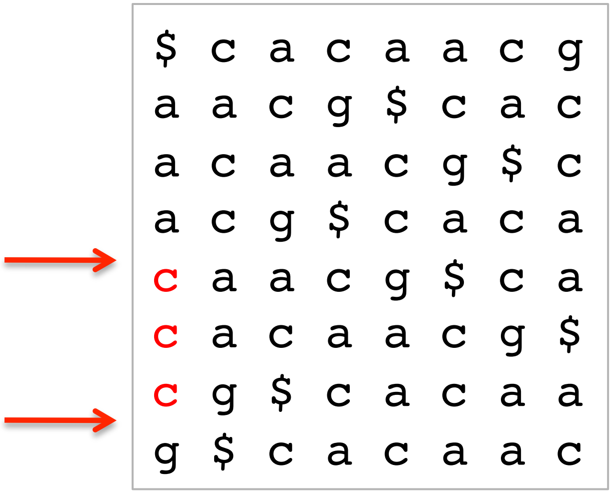
Lookup AAC

```
$ c a c a a c g
a a c g $ c a c
a c a a c g $ c
a c g $ c a c a
c a a c g $ c a
c a c a a c g $
c g $ c a c a a
g $ c a c a a c
```

```
Range ← range of last character in 1st column
While characters left (and nonzero range):
    Lookup first and last match to preceding character in final column
    Range ← LF-mapping of first and last match
```

Source: Mott et. al., Wellcome Trust

Lookup AAC

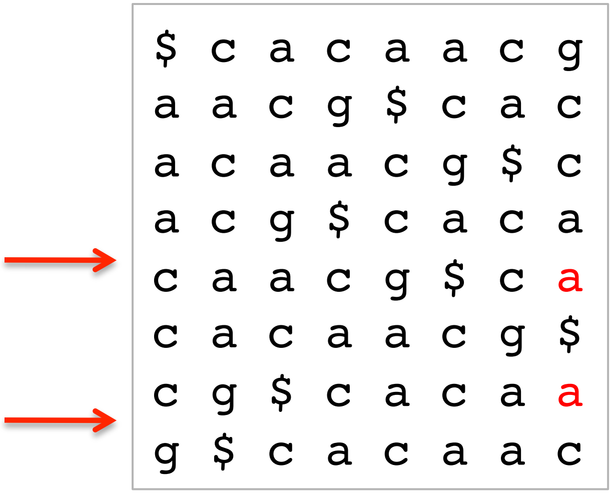


$	c	a	c	a	a	c	g
a	a	c	g	$	c	a	c
a	c	a	a	c	g	$	c
a	c	g	$	c	a	c	a
c	a	a	c	g	$	c	a
c	a	c	a	a	c	g	$
c	g	$	c	a	c	a	a
g	$	c	a	c	a	a	c

```
Range ← range of last character in 1$^{st}$ column
While characters left (and nonzero range):
    Lookup first and last match to preceding character in final column
    Range ← LF-mapping of first and last match
```

Source: Mott et. al., Wellcome Trust

Lookup AAC

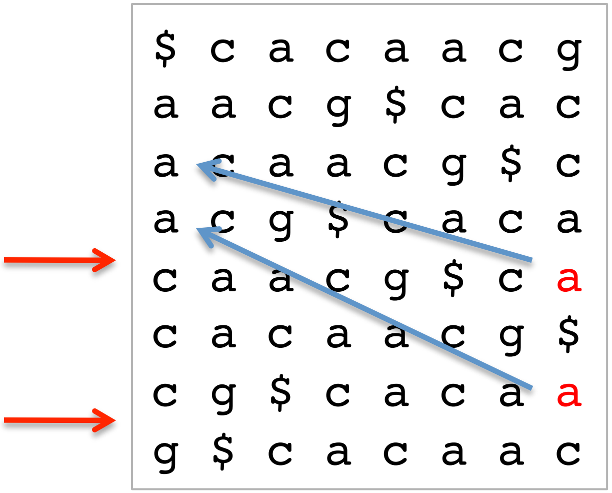


```
$ c a c a a c g
a a c g $ c a c
a c a a c g $ c
a c g $ c a c a
c a a c g $ c a
c a c a a c g $
c g $ c a c a a
g $ c a c a a c
```

```
Range ← range of last character in 1st column
While characters left (and nonzero range):
    Lookup first and last match to preceding character in final column
    Range ← LF-mapping of first and last match
```

Source: Mott et. al., Wellcome Trust

Lookup AAC

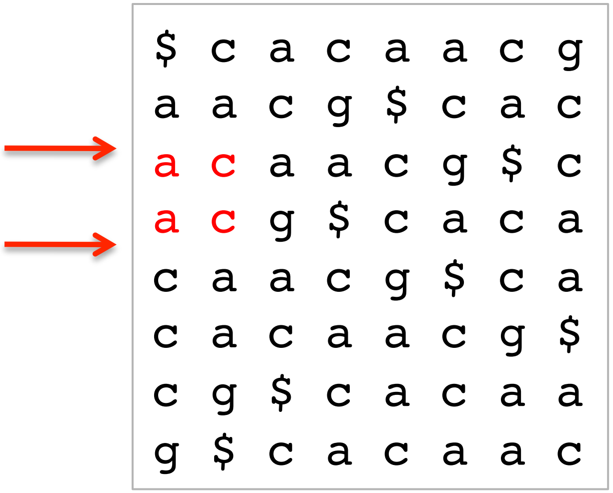


$	c	a	c	a	a	c	g
a	a	c	g	$	c	a	c
a	c	a	a	c	g	$	c
a	c	g	$	c	a	c	a
c	a	a	c	g	$	c	a
c	a	c	a	a	c	g	$
c	g	$	c	a	c	a	a
g	$	c	a	c	a	a	c

```
Range ← range of last character in 1st column
While characters left (and nonzero range):
    Lookup first and last match to preceding character in final column
    Range ← LF-mapping of first and last match
```

Source: Mott et. al., Wellcome Trust

Lookup AAC

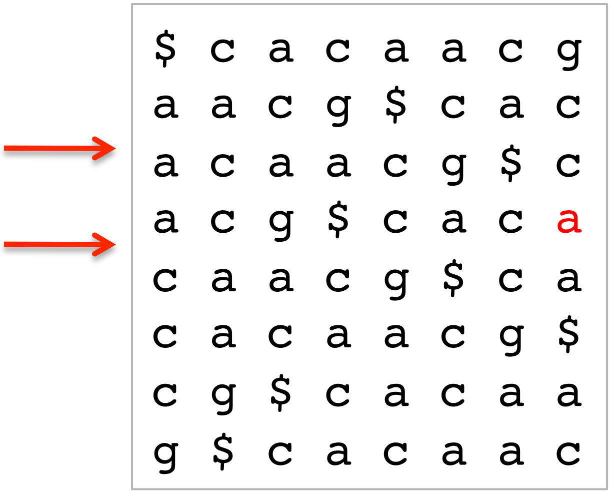


$	c	a	c	a	a	c	g
a	a	c	g	$	c	a	c
a	c	a	a	c	g	$	c
a	c	g	$	c	a	c	a
c	a	a	c	g	$	c	a
c	a	c	a	a	c	g	$
c	g	$	c	a	c	a	a
g	$	c	a	c	a	a	c

```
Range ← range of last character in 1st column
While characters left (and nonzero range):
    Lookup first and last match to preceding character in final column
    Range ← LF-mapping of first and last match
```

Source: Mott et. al., Wellcome Trust

Lookup AAC

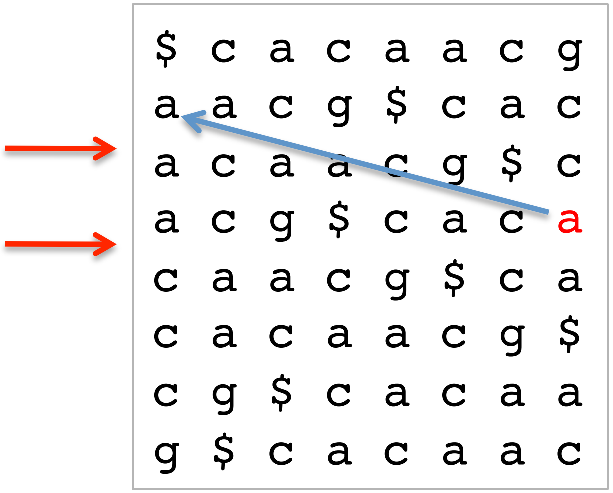


$	c	a	c	a	a	c	g
a	a	c	g	$	c	a	c
a	c	a	a	c	g	$	c
a	c	g	$	c	a	c	a
c	a	a	c	g	$	c	a
c	a	c	a	a	c	g	$
c	g	$	c	a	c	a	a
g	$	c	a	c	a	a	c

```
Range ← range of last character in 1st column
While characters left (and nonzero range):
    Lookup first and last match to preceding character in final column
    Range ← LF-mapping of first and last match
```

Source: Mott et. al., Wellcome Trust

Lookup AAC

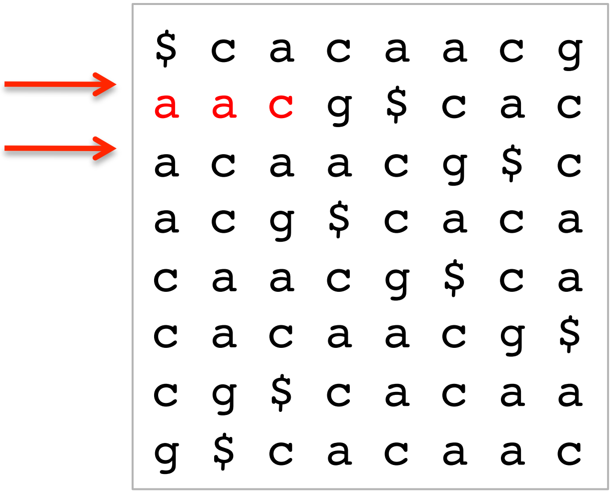


$	c	a	c	a	a	c	g
a	a	c	g	$	c	a	c
a	c	a	a	c	g	$	c
a	c	g	$	c	a	c	a
c	a	a	c	g	$	c	a
c	a	c	a	a	c	g	$
c	g	$	c	a	c	a	a
g	$	c	a	c	a	a	c

```
Range ← range of last character in 1st column
While characters left (and nonzero range):
    Lookup first and last match to preceding character in final column
    Range ← LF-mapping of first and last match
```

Source: Mott et. al., Wellcome Trust

Lookup AAC



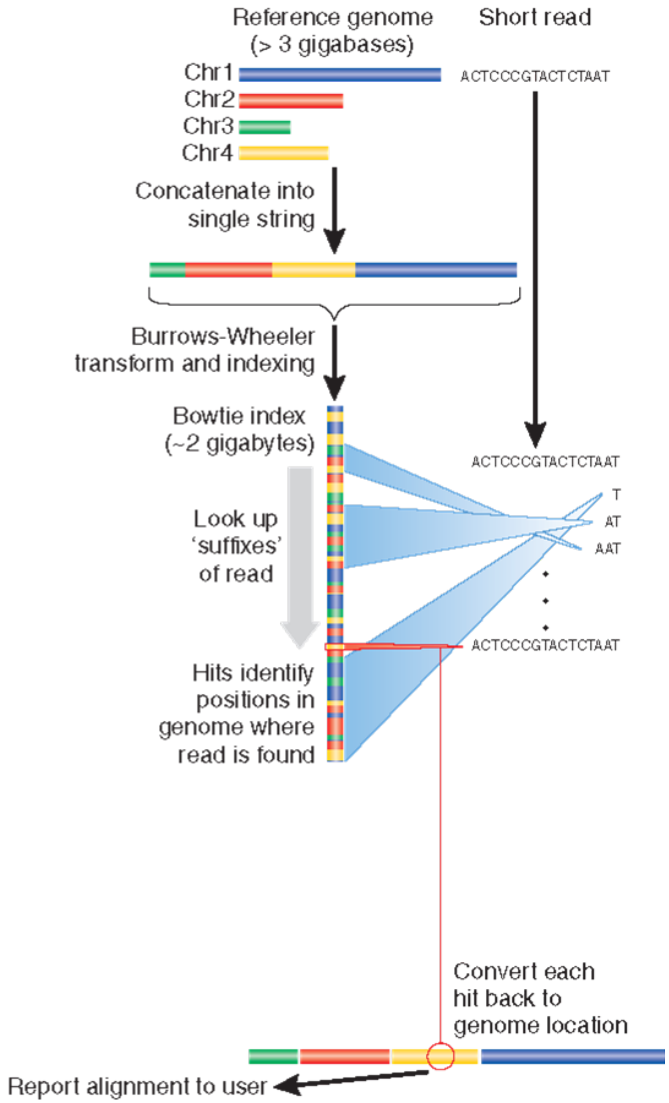
$	c	a	c	a	a	c	g
a	a	c	g	$	c	a	c
a	c	a	a	c	g	$	c
a	c	g	$	c	a	c	a
c	a	a	c	g	$	c	a
c	a	c	a	a	c	g	$
c	g	$	c	a	c	a	a
g	$	c	a	c	a	a	c

```
Range ← range of last character in 1st column
While characters left (and nonzero range):
    Lookup first and last match to preceding character in final column
    Range ← LF-mapping of first and last match
```

Source: Mott et. al., Wellcome Trust

However there is a *great* deal more to an actual aligner

- This is a workflow of BOWTIE
  - Ultrafast and memory-efficient alignment of short DNA sequences to the human genome, Ben Langmead, Cole Trapnell, Mihai Pop and Steven L Salzberg, Genome Biology 2009, 10:R25
  - Indexes the genome, not the reads, relatively low memory usage for BW, but single threaded.
- But there are many more implementations as we will see …

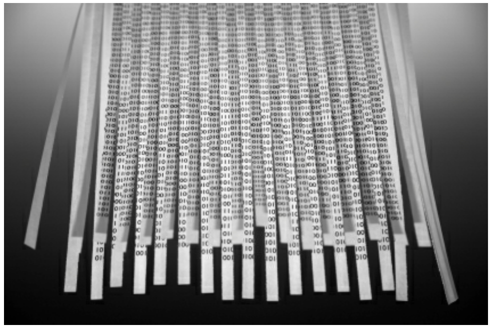

Reference genome
(> 3 gigabases)
Short read
Chr1
Chr2
Chr3
Chr4
ACTCCCGTACTCTAAT
Concatenate into single string
Burrows-Wheeler transform and indexing
Bowtie index
(~2 gigabytes)
ACTCCCGTACTCTAAT
T
AT
AAT
Look up 'suffixes' of read
ACTCCCGTACTCTAAT
Hits identify positions in genome where read is found
Convert each hit back to genome location
Report alignment to user

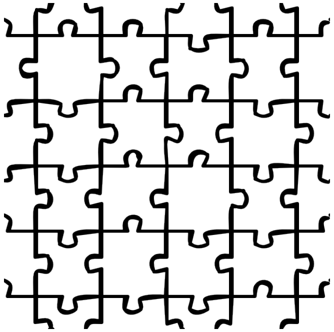
Adapted from Michael Stromberg, Illumina, Canadian Bioinformatics Workshops 2011 pause

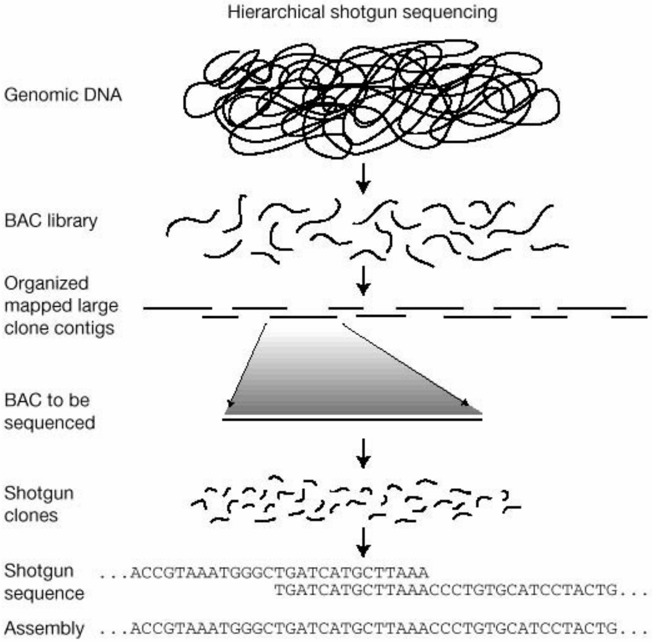
Assembly

- If you already have a reference genome against which to align … pick an aligner/ mapper
- If you don't, then you are faced with de-Novo assembly
- Two general types of assemblers
  - Overlap Consensus
  - De Bruijn Graphs
- For de-Novo assembly from HT sequencing de Bruijn methods have proven most efficient

Put that back together →

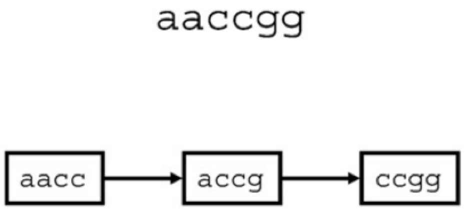





Hierarchical shotgun sequencing
Genomic DNA
BAC library
Organized mapped large clone contigs
BAC to be sequenced
Shotgun clones
Shotgun sequence
...ACCGTAAATGGGCTGATCATGCTTAAA
TGATCATGCTTAAACCCTGTGCATCCTACTG...
Assembly
...ACCGTAAATGGGCTGATCATGCTTAAACCCTGTGCATCCTACTG...

De Bruijn Graph assembly

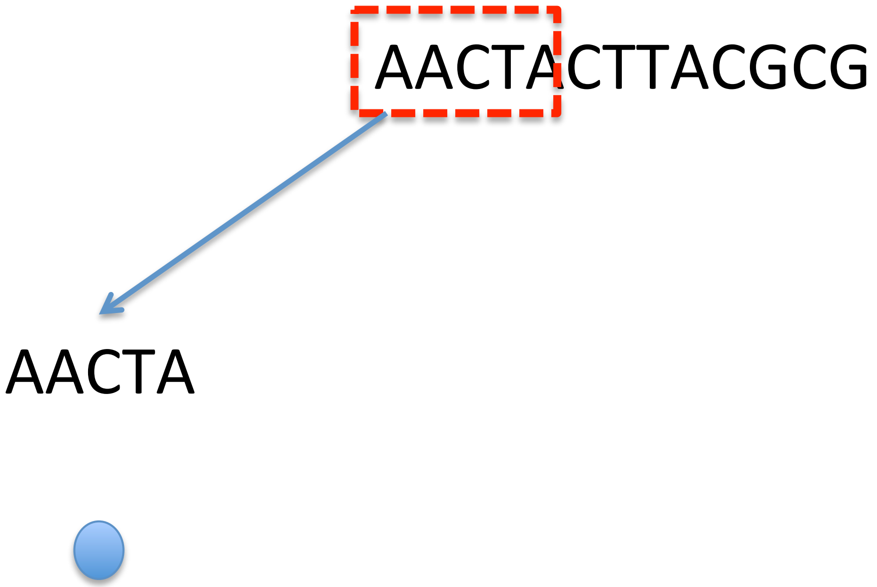
- **Wikipedia: “De Bruijn graph** of *m* symbols is a directed graph representing overlaps between sequences of symbols”
- Use as an assembler - Concept:
  - Decompose all reads into k-mers (words of fixed length k)
  - Construct a graph with the k-mers as vertices and the directed edges are the connections between a k-mers that overlap by k-1
  - The assembly is the path through the graph



The de Bruijn Graph

a representation of all possible paths joining reads together
Pevsner, PNAS 2001

Choose a word length k (5 in this example, but larger in applications)

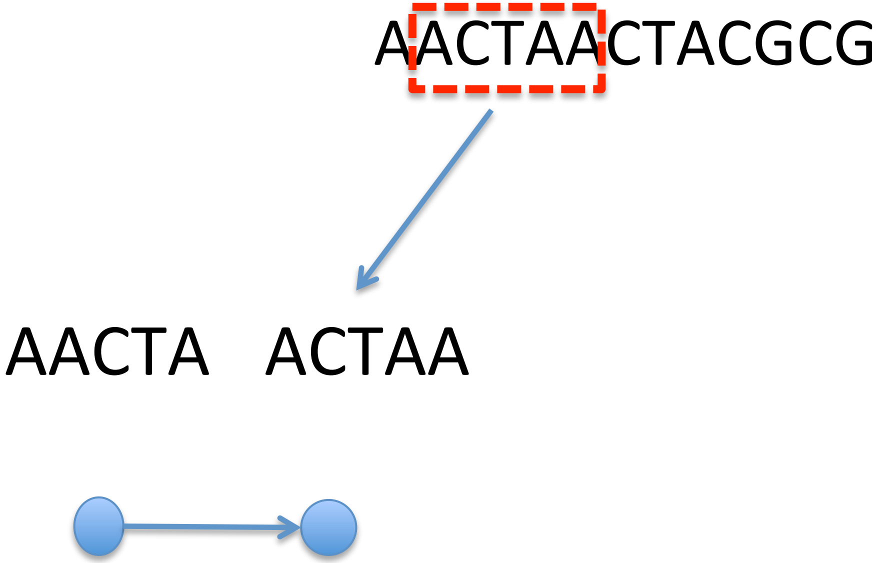


AACTACTTACGCG

AACTA

Source: Mott et. al., Wellcome Trust

The de Bruijn Graph

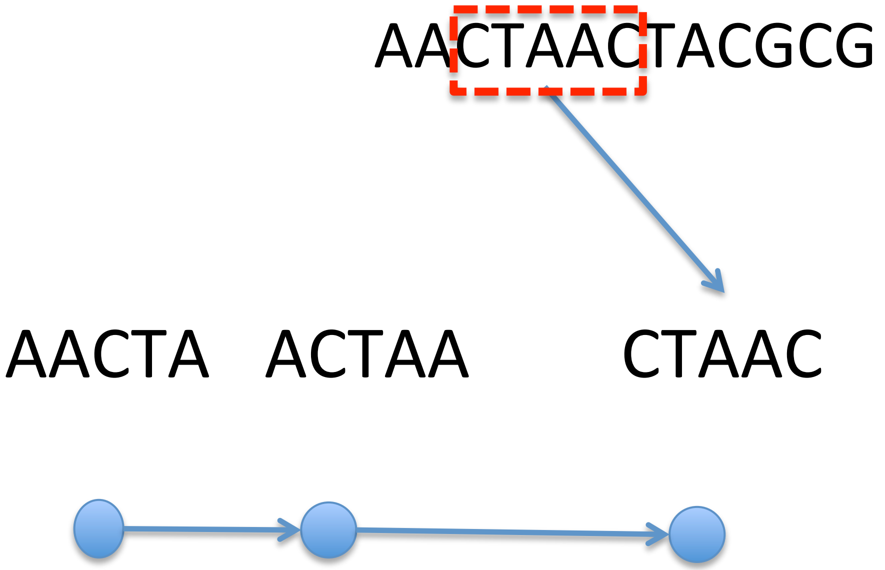


AACTAACTACGCG

AACTA ACTAA

Source: Mott et. al., Wellcome Trust

The de Bruijn Graph

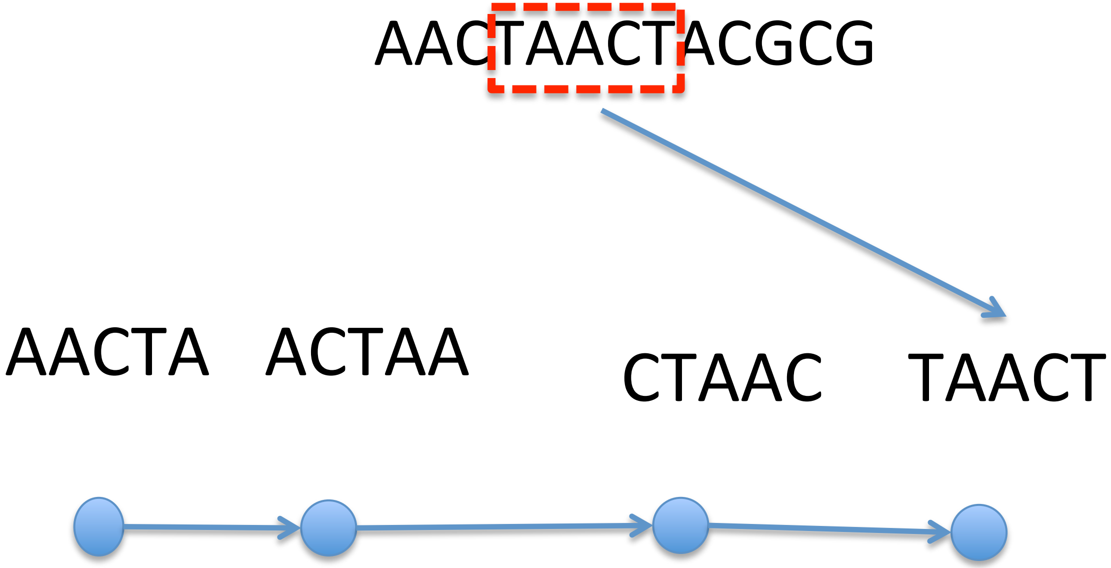


AACTAACTACGCG

AACTA ACTAA CTAAC

Source: Mott et. al., Wellcome Trust

The de Bruijn Graph

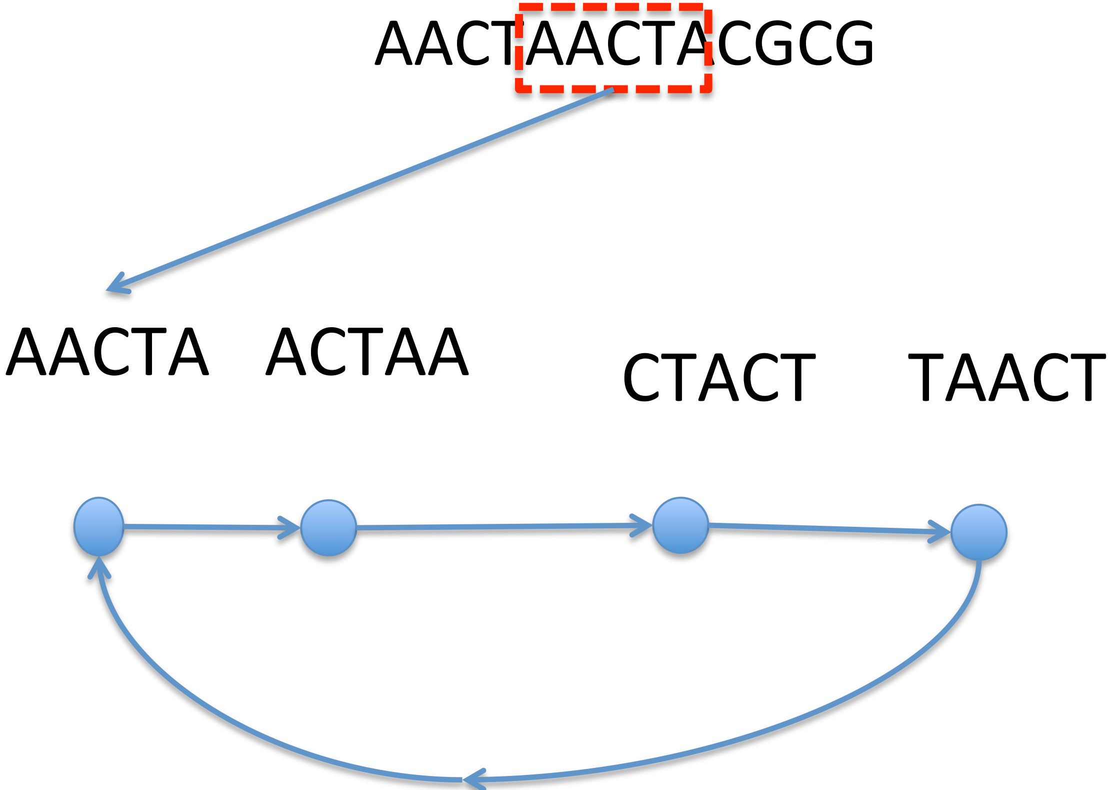


AACTAACTACGCG

AACTA ACTAA CTAAC TAACT

Source: Mott et. al., Wellcome Trust

The de Bruijn Graph



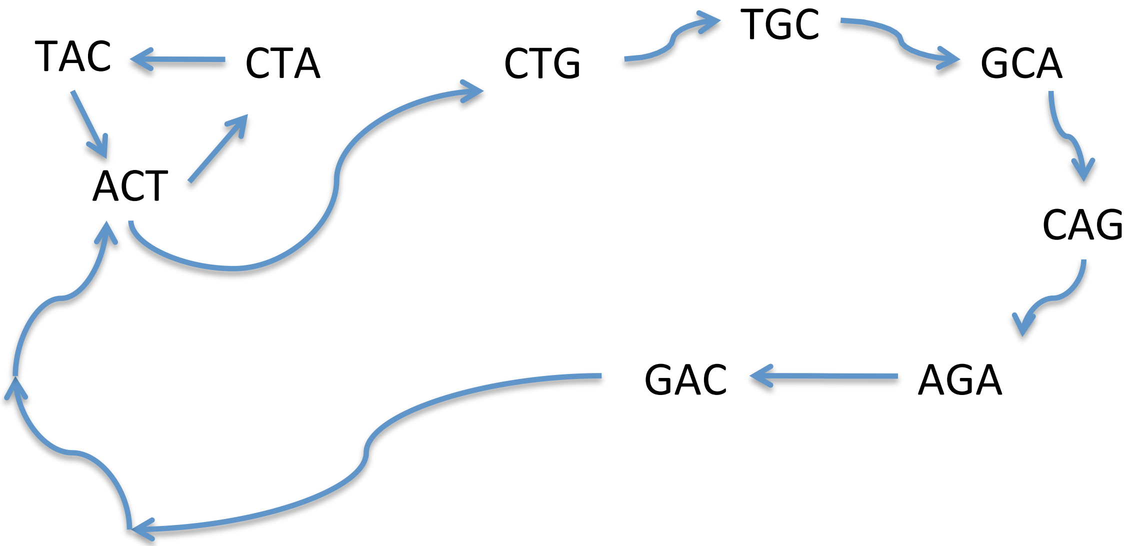
AACTAACTACGCG

AACTA ACTAA CTACT TAACT

Source: Mott et. al., Wellcome Trust

Same sequence, different k=3

ACTACTACTGCAGACTACT



TAC

CTA

CTG

TGC

GCA

ACT

CAG

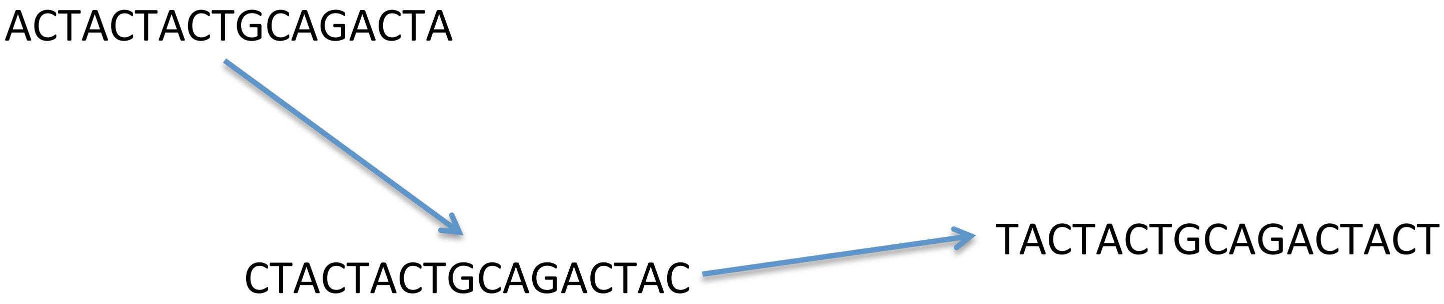
GAC

AGA

Source: Mott et. al., Wellcome Trust

Same sequence, different k=17

ACTACTACTGCAGACTACT



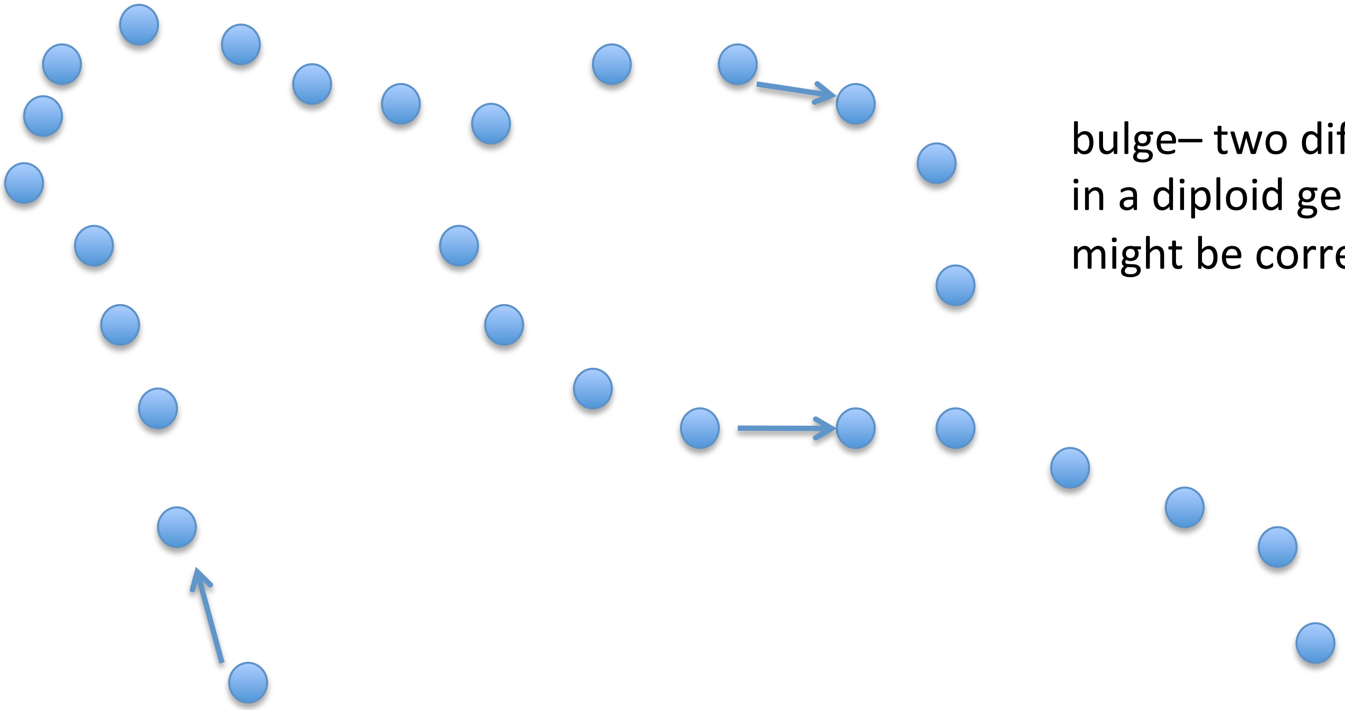
ACTACTACTGCAGACTA

CTACTACTGCAGACTAC

TACTACTGCAGACTACT

Source: Mott et. al., Wellcome Trust

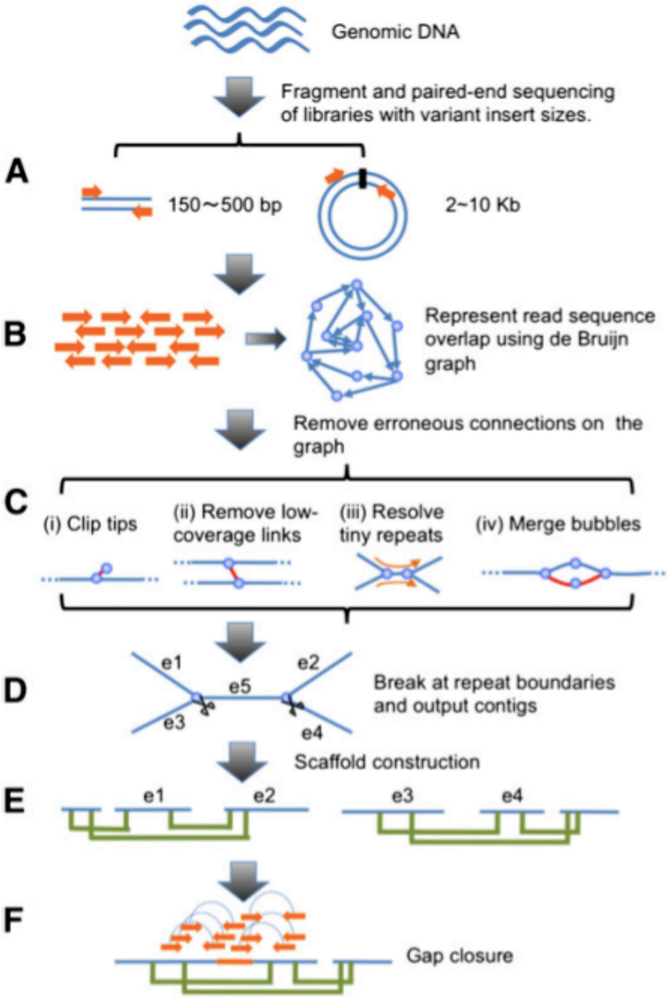
Recovering unambiguous contigs



bulge– two different paths;
in a diploid genome both
might be correct

Source: Mott et. al., Wellcome Trust

But there is a lot more than “just” the de Bruijn Graph


Genomic DNA
Fragment and paired-end sequencing of libraries with variant insert sizes.
A
150~500 bp
2~10 Kb
B
Represent read sequence overlap using de Bruijn graph
Remove erroneous connections on the graph
C
(i) Clip tips
(ii) Remove low-coverage links
(iii) Resolve tiny repeats
(iv) Merge bubbles
D
e1
e2
e5
e3
e4
Break at repeat boundaries and output contigs
Scaffold construction
E
e1
e2
e3
e4
F
Gap closure

This is the workflow of SOAPdenovo from:

De novo assembly of human genomes with massively parallel short read sequencing, Ruiqiang Li, et. al., Genome Res, 2010 20: 265-272

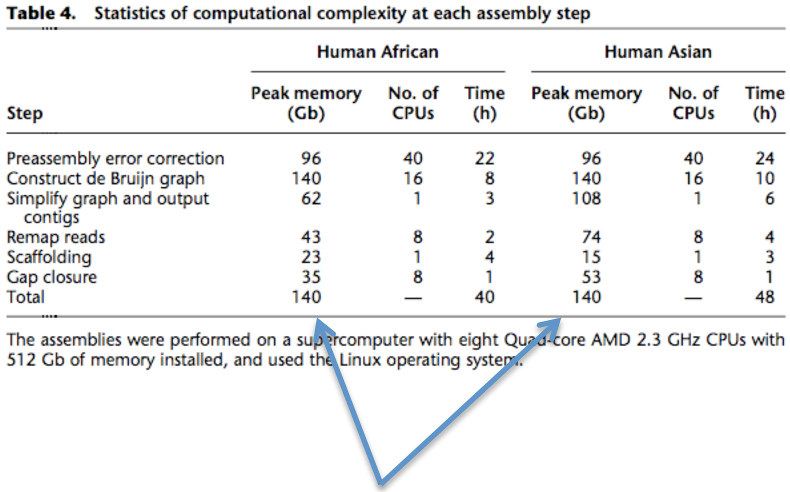


Table 4. Statistics of computational complexity at each assembly step

Step	Human African			Human Asian		
	Peak memory (Gb)	No. of CPUs	Time (h)	Peak memory (Gb)	No. of CPUs	Time (h)
Preassembly error correction	96	40	22	96	40	24
Construct de Bruijn graph	140	16	8	140	16	10
Simplify graph and output contigs	62	1	3	108	1	6
Remap reads	43	8	2	74	8	4
Scaffolding	23	1	4	15	1	3
Gap closure	35	8	1	53	8	1
Total	140	—	40	140	—	48

The assemblies were performed on a supercomputer with eight Quad-core AMD 2.3 GHz CPUs with 512 Gb of memory installed, and used the Linux operating system.

Remember these numbers for later

Source: Mott et. al., Wellcome Trust

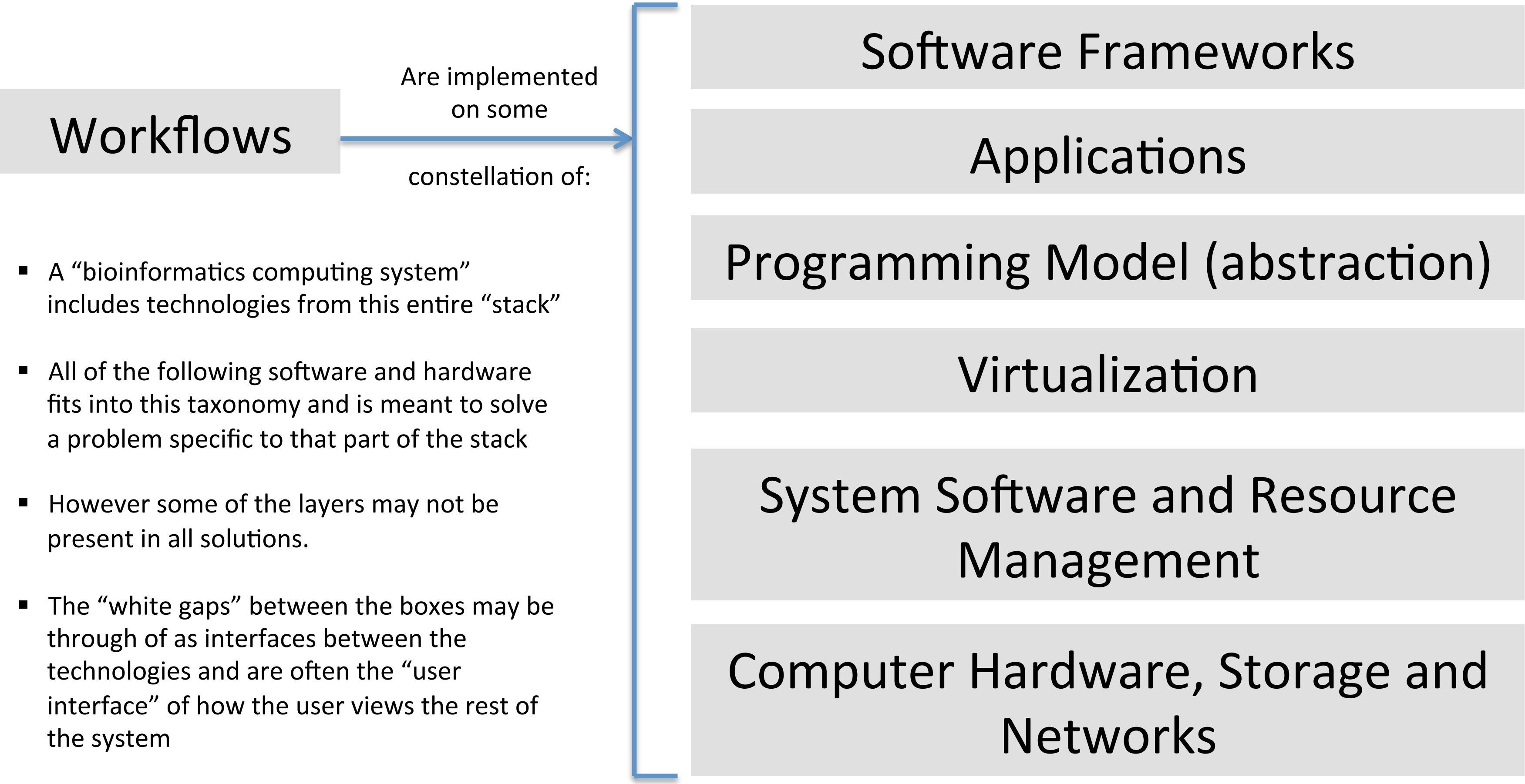
Break

After the break, let's look at software and hardware systems and the computing challenges they are meant to address

Now that we know a bit about some important algorithms

- Let's look at the the systems and infrastructures that are out there
  - The aspects of the computing demands that they address
  - These are generally the ways that you will access the infrastructures
- Portals, Web interfaces, tool systems
  - local, departmental, big data center and cloud
- Accelerators, big memory machines, accelerator-appliances with back -ends

First some Taxonomy



Workflows

Are implemented on some constellation of:

Software Frameworks

Applications

Programming Model (abstraction)

Virtualization

System Software and Resource Management

Computer Hardware, Storage and Networks

- A “bioinformatics computing system” includes technologies from this entire “stack”
- All of the following software and hardware fits into this taxonomy and is meant to solve a problem specific to that part of the stack
- However some of the layers may not be present in all solutions.
- The “white gaps” between the boxes may be through of as interfaces between the technologies and are often the “user interface” of how the user views the rest of the system
- Keep this in mind as we discuss the following software and hardware

Summary of computing barriers

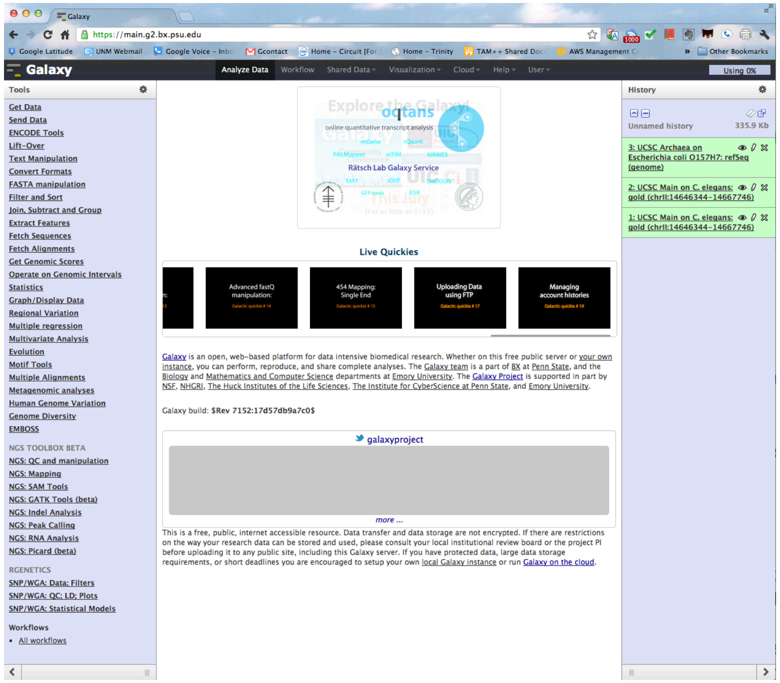
- many (most) applications are single threaded.
- many (most) applications are written for a single address space.
  - NGS-size data quickly pushes 1) and 2) beyond the capacity of a single node
    - Need multiple threads
    - A large memory footprint
- Some algorithms (SW as an example) scale quadraticlly with the size of the problem
  - Motivating algorithmic substitution or hardware acceleration
- Working subsets are growing too large to fit into available memory
  - Mapping/aligning with BW and assembly with De Bruijn are good examples
  - Motivating algorithmic innovations and novel approaches to large memory computers.
- The amount of data barely fits into currently available disk space. (And soon might not – see the first part of the talk)
- Databases are distributed and will likely stay that way
  - Motivating much talk of "bringing the computing to the data"
  - Of preprocessing for downstream upload, etc….
  - You will see several ideas for solutions …

Software Frameworks

- Available as a “tar ball” for your local (or cloud) installation pleasure
  - Galaxy
  - Bioconductor, R
  - …
- Presented as a portal and backed by some computing horsepower
  - Galaxy
  - Bionimbus
  - IMG/…
  - MG-RAST
  - …
- Or available as a cloud image, launch-able on your cloud (Eucalyptus, OpenStack)or on a provider (EC2, Google, etc.)
  - Galaxy
  - Bionimbus
  - Bioconductor, R
  - CloudBioLinux, descended from BioLinux
  - …

Galaxy

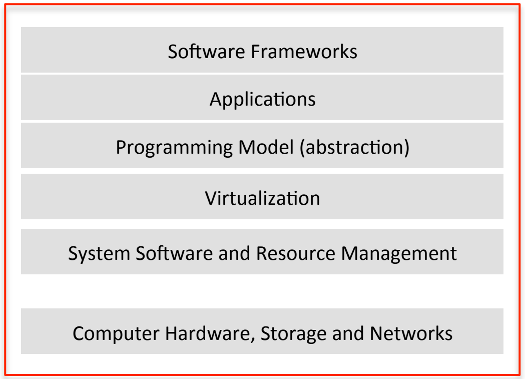
- A web based platform for analysis of large genomic datasets (galaxy.psu.edu)
- Integrates many tools within one interface:
  - Including a “Next Generation Sequencing Toolbox”
  - And “access” to many databases (data must be loaded on request)
- Available through a PSU.edu portal, for tar-ball download, or as an EC2 image



Galaxy is an open, web-based platform for data intensive biomedical research. Whether on this free public server or your own instance, you can perform, reproduce, and share complete analyses. The Galaxy team is a part of BX at Penn State, and the Biology and Mathematics and Computer Science departments at Emory University. The Galaxy Project is supported in part by NSF, NHGRI, The Huck Institutes of the Life Sciences, The Institute for CyberScience at Penn State, and Emory University.

Galaxy build: **$Rev 7152:17d57db9a7c0$**

This is a free, public, internet accessible resource. Data transfer and data storage are not encrypted. If there are restrictions on the way your research data can be stored and used, please consult your local institutional review board or the project PI before uploading it to any public site, including this Galaxy server. If you have protected data, large data storage requirements, or short deadlines you are encouraged to setup your own local Galaxy instance or run Galaxy on the cloud.


Software Frameworks

Applications

Programming Model (abstraction)

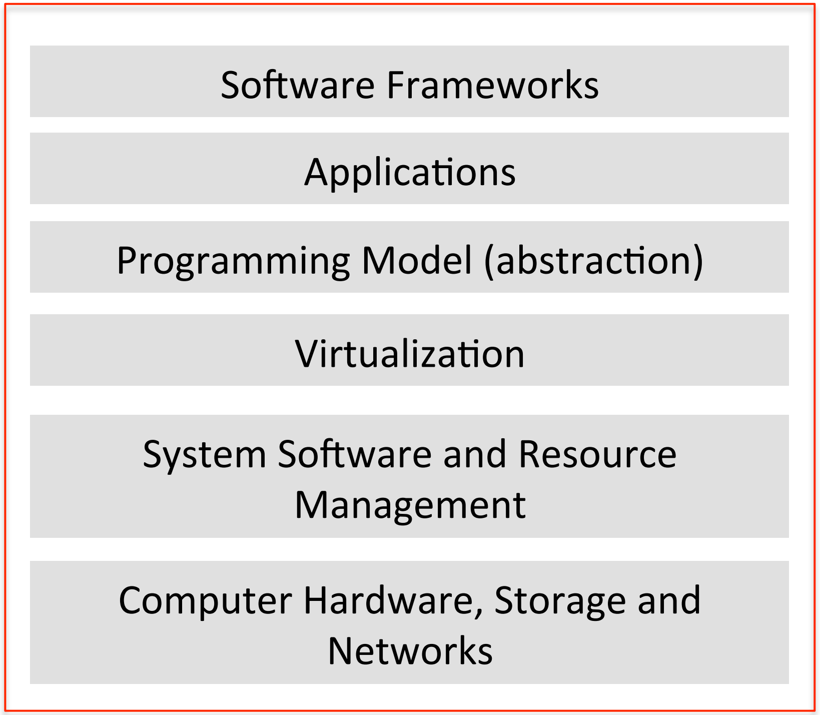
Virtualization

System Software and Resource Management

Computer Hardware, Storage and Networks

Other Framework Examples

- Bioconductor (http://www.bioconductor.org/)
  - U. Wash
  - Built on top of "R"
  - Tar-ball or EC2 image
- Bionimbus (http://www.bionimbus.org/)
  - U. Chicago
  - Portal (http://bc.bionimbus.org/Bionimbus/), Tar-Ball, C2 or private cloud images
- CloudBioLinux
  - JCVI, NEBC Bioinformatics Centre, Harvard, Galaxy
  - Was tar-ball BioLinux, but now available as CloudBioLinux in an Amazon EC2 image


Software Frameworks
Applications
Programming Model (abstraction)
Virtualization
System Software and Resource Management
Computer Hardware, Storage and Networks

Bioconductor

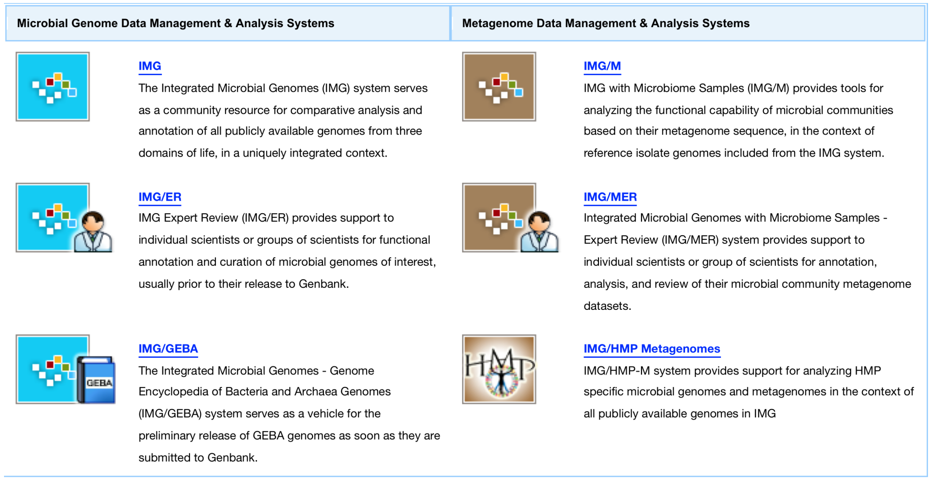
- Bioconductor provides tools for the analysis and comprehension of high-throughput genomic data. Bioconductor uses the R statistical programming language, and is open source and open development. It has two releases each year, 554 software packages, and an active user community. Bioconductor is also available as an Amazon Machine Image (AMI).
- http://www.bioconductor.org/
- Bioconductor is built on top of "R" a statistics analysis engine
  - R is a language and environment for statistical computing and graphics. It is a GNU project which is similar to the S language and environment which was developed at Bell Laboratories. There are some important differences, but much code written for S runs unaltered under R.
  - R provides a wide variety of statistical (linear and nonlinear modeling, classical statistical tests, time-series analysis, classification, clustering, ...) and graphical techniques, and is highly extensible.
  - http://www.r-project.org/

Bionimbus

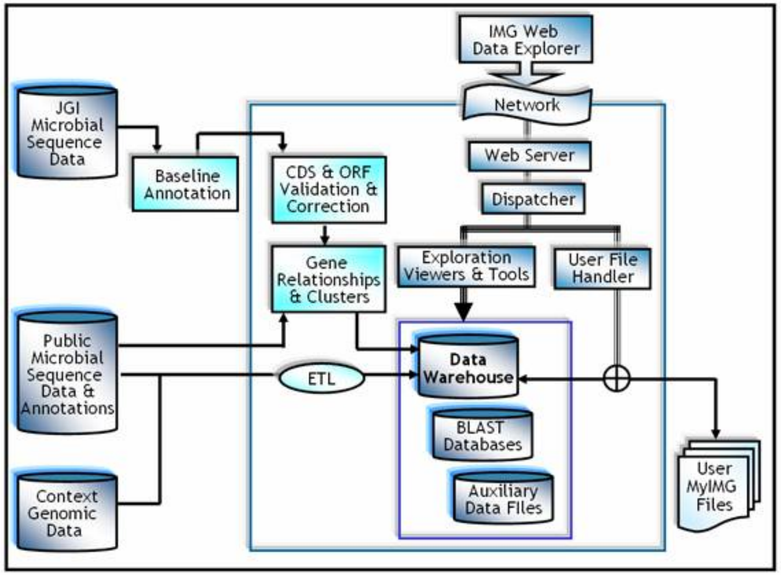
- **Overview.** Bionimbus is an open source cloud-based system for managing, analyzing and sharing genomic data that has been developed by the Institute for Genomics and Systems Biology (IGSB) at the University of Chicago. Bionimbus is designed to support next-generation sequencing instruments and integrates technology for the analyzing and transporting large datasets. There is an open source version of Bionimbus available to those who wish to set up their own clouds. There is also a Bionimbus Community Cloud operated by the 501(c)(3) Open Cloud Consortium's Open Science Data Cloud that includes a variety of public genomics and related data.
- **Bionimbus Community Cloud.** There is a Bionimbus Community Cloud, which research collaborators can log into and use.
- Bionimbus uses the Open Cloud Consortium's Open Science Data Cloud (OSDC) for its infrastructure. The first generation of Bionimbus used 7 racks of equipment containing approximately 3000 cores and 1 PB of disk that Yahoo! donated to the OSDC. Currently, we are using approximately six racks of equipment that the Gordon and Betty Moore Foundation has funded. Cisco has provided access to the Cisco C-Wave so that we can connect the four OSDC data centers together with 10 Gbps wide area networks.
- **Virtual machine images.** We develop and maintain Bionimbus machine images that can be run on:
- i) The Bionimbus Community Cloud.
- ii) Public clouds such as Amazon's.
- iii) Your own Eucalyptus or OpenStack-based private clouds.
- **Your own Bionimbus cloud**. The Bionimbus system itself is open source and you can build your own private Bionimbus clouds.

Web only Frameworks: The IMG Web Tools

- Oriented around microbial and metagenomic analysis
- Available only through the Joint Genome Institute's web portal
  - Consolidates applications, databases and computation
- Data Center
  - "Our data storage system consists of a couple Oracle databases and file systems. We also have several backend pipelines that use the supercomputing facility in NERSC to perform gene calling and annotations."



Microbial Genome Data Management & Analysis Systems	Metagenome Data Management & Analysis Systems
IMG The Integrated Microbial Genomes (IMG) system serves as a community resource for comparative analysis and annotation of all publicly available genomes from three domains of life, in a uniquely integrated context.	**IMG/M** IMG with Microbiome Samples (IMG/M) provides tools for analyzing the functional capability of microbial communities based on their metagenome sequence, in the context of reference isolate genomes included from the IMG system.
IMG/ER IMG Expert Review (IMG/ER) provides support to individual scientists or groups of scientists for functional annotation and curation of microbial genomes of interest, usually prior to their release to Genbank.	**IMG/MER** Integrated Microbial Genomes with Microbiome Samples - Expert Review (IMG/MER) system provides support to individual scientists or group of scientists for annotation, analysis, and review of their microbial community metagenome datasets.
IMG/GEBA The Integrated Microbial Genomes - Genome Encyclopedia of Bacteria and Archaea Genomes (IMG/GEBA) system serves as a vehicle for the preliminary release of GEBA genomes as soon as they are submitted to Genbank.	**IMG/HMP Metagenomes** IMG/HMP-M system provides support for analyzing HMP specific microbial genomes and metagenomes in the context of all publicly available genomes in IMG



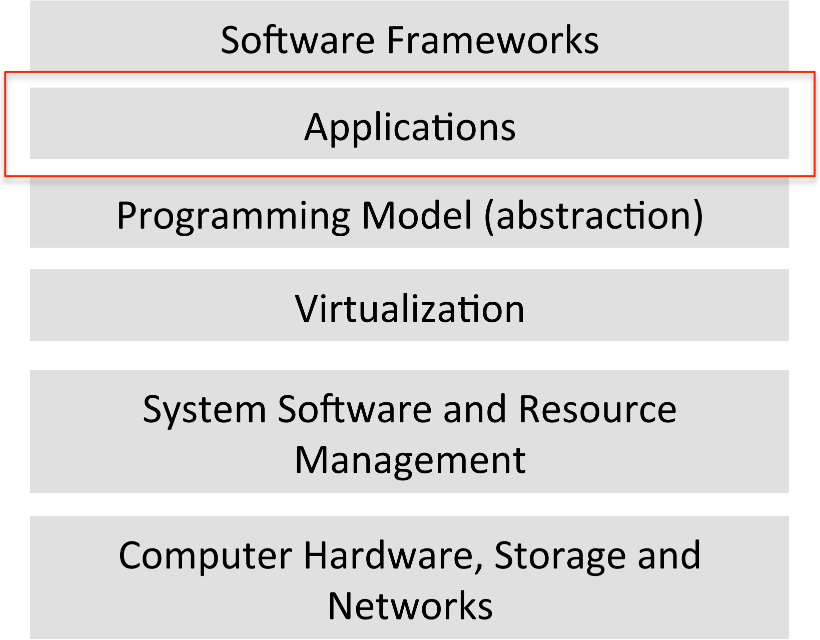
More Web-only Frameworks

MG-RAST
metagenomics analysis server

- Argonne National Laboratory
- Annotation and comparative analysis of metagenomes,
  - currently prokaryotes and viruses
- Databases
  - integrated into the M5NR non-redundant database using the M5NR tools.
- Protein databases:
  - The SEED, GenBank, RefSeq, IMG/M, UniProt, eggNOGG, KEGG, PATRIC,
- Ribosomal RNA databases:
  - Greengenes, SILVA, RDP
- Bioinformatics Tools:
  - FragGeneScan,BLAT , QIIME , Biopython, Bowtie ,sff_extract, Dynamic Trim, Krona , Raphaël JavaScript Library ,jQuery, Circos cURL
- Behind the scenes:
  - Perl, Python, R, Google's V8 JavaScript engine, Node.js, JumpLoader, Nginx, OpenStack
- Compute infrastructure?
  - note that MG-RAST is at Argonne, also one of the pre-eminent HPC laboratories

Applications that implement the algorithms

- There is an absolute *wealth* of available competing software
- Each with variations specific
  - to the domain
  - to the computing challenge
- However the core applications fit into the taxonomy like this:
- Most of the time the apps are embedded in a framework, but are sometimes directly available "at the command line."
- They are always supported by the stack technologies below them, whether delivered by laptop, workstation, server or cloud.
- The following slides give a whirlwind tour of apps



Software Frameworks

Applications

Programming Model (abstraction)

Virtualization

System Software and Resource Management

Computer Hardware, Storage and Networks

Mappers/Aligners

- BOWTIE - Ultrafast, memory-efficient short read aligner. It aligns short DNA sequences (reads) to the human genome at a rate of 25 million reads per hour on a typical workstation with 2 gigabytes of memory. Uses a Burrows-Wheeler-Transformed (BWT) index. Link to discussion thread here. Written by Ben Langmead and Cole Trapnell. Linux, Windows, and Mac OS X.
- MAQ - Ultrafast, memory-efficient short read aligner. It aligns short DNA sequences (reads) to the human genome at a rate of 25 million reads per hour on a typical workstation with 2 gigabytes of memory. Uses a Burrows-Wheeler-Transformed (BWT) index. Link to discussion thread here. Written by Ben Langmead and Cole Trapnell. Linux, Windows, and Mac OS X.
- inGAP - Integrated Next-gen Genome Analysis Platform
- DNA star - DNASTAR has software solutions to simplify your reference-guided assembly projects using Next-Generation Roche 454 Life Sciences, Illumina, ABI SOLiD, Helicos and/or Sanger data.
- BWA - Heng Lee's BWT Alignment program - a progression from Maq. BWA is a fast light-weighted tool that aligns short sequences to a sequence database, such as the human reference genome. By default, BWA finds an alignment within edit distance 2 to the query sequence. C++ source.
- GenomeMapper - GenomeMapper is a short read mapping tool designed for accurate read alignments. It quickly aligns millions of reads either with ungapped or gapped alignments. A tool created by the 1001 Genomes project. Source for POSIX.
- Genomic Next-generation Universal MAPper (gnumap) - The Genomic Next-generation Universal MAPper (gnumap) is a program designed to accurately map sequence data obtained from next-generation sequencing machines (specifically that of Solexa/Illumina) back to a genome of any size. It seeks to align reads from nonunique repeats using statistics. From authors at Brigham Young University. C source/Unix.
- RMAP - Assembles 20 - 64 bp Illumina reads to a FASTA reference genome. By Andrew D. Smith and Zhenyu Xuan at CSHL. (published in BMC Bioinformatics). POSIX OS required.

More Mappers/Aligners

- MOSAIK - MOSAIK produces gapped alignments using the Smith-Waterman algorithm. Features a number of support tools. Support for Roche FLX, Illumina, SOLiD, and Helicos. Written by Michael Str
- mr & mrs FAST - mrFAST & mrsFAST are designed to map short reads generated with the Illumina platform to reference genome assemblies; in a fast and memory-efficient manner. Robust to INDELs and MrsFAST has a bisulphite mode. Authors are from the University of Washington. C as source.
- MUMmer- MUMmer is a modular system for the rapid whole genome alignment of finished or draft sequence. Released as a package providing an efficient suffix tree library, seed-and-extend alignment, SNP detection, repeat detection, and visualization tools. Version 3.0 was developed by Stefan Kurtz, Adam Phillippy, Arthur L Delcher, Michael Smoot, Martin Shumway, Corina Antonescu and Steven L Salzberg - most of whom are at The Institute for Genomic Research in Maryland, USA. POSIX OS required.
- NOVOCRAFT- Tools for reference alignment of paired-end and single-end Illumina reads. Uses a Needleman-Wunsch algorithm. Can support Bis-Seq. Commercial. Available free for evaluation, educational use and for use on open not-for-profit projects. Requires Linux or Mac OS X.
- PASS- I t supports Illumina, SOLiD and Roche-FLX data formats and allows the user to modulate very finely the sensitivity of the alignments. Spaced seed intial filter, then NW dynamic algorithm to a SW(like) local alignment. Authors are from CRIBI in Italy. Win/Linux.
- SOAP- SOAP (Short Oligonucleotide Alignment Program). A program for efficient gapped and ungapped alignment of short oligonucleotides onto reference sequences. The updated version uses a BWT. Can call SNPs and INDELs. Author is Ruiqiang Li at the Beijing Genomics Institute. C++, POSIX.

Mapping Assembly

- ZOOM- ZOOM (Zillions Of Oligos Mapped) is designed to map millions of short reads, emerged by next-generation sequencing technology, back to the reference genomes, and carry out post-analysis. ZOOM is developed to be highly accurate, flexible, and user-friendly with speed being a critical priority. Commercial. Supports Illumina and SOLiD data.
- SOCS- Aligns SOLiD data. SOCS is built on an iterative variation of the Rabin-Karp string search algorithm, which uses hashing to reduce the set of possible matches, drastically increasing search speed. Authors are Ondov B, Varadarajan A, Passalacqua KD and Bergman NH.
- SHRiMP- Assembles to a reference sequence. Developed with Applied Biosystem's colourspace genomic representation in mind. Authors are Michael Brudno and Stephen Rumble at the University of Toronto. POSIX.
- SLIDER - An application for the Illumina Sequence Analyzer output that uses the probability files instead of the sequence files as an input for alignment to a reference sequence or a set of reference sequences. Authors are from BCGSC.
- BFAST - Blat-like Fast Accurate Search Tool. Written by Nils Homer, Stanley F. Nelson and Barry Merriman at UCLA.
- ELAND - Efficient Large-Scale Alignment of Nucleotide Databases. Whole genome alignments to a reference genome. Written by Illumina author Anthony J. Cox for the Solexa 1G machine.
- Exonerate - Various forms of pairwise alignment (including Smith-Waterman-Gotoh) of DNA/protein against a reference. Authors are Guy St C Slater and Ewan Birney from EMBL. C for POSIX.
- GMAP - GMAP (Genomic Mapping and Alignment Program) for mRNA and EST Sequences. Developed by Thomas Wu and Colin Watanabe at Genentec. C/Perl for Unix.

Mapping Assembly

- GNUMAP- The Genomic Next-generation Universal MAPper (gnumap) is a program designed to accurately map sequence data obtained from next-generation sequencing machines (specifically that of Solexa/Illumina) back to a genome of any size. It seeks to align reads from nonunique repeats using statistics. From authors at Brigham Young University. C source/Unix.
- SeqMap - Supports up to 5 or more bp mismatches/ INDELs. Highly tunable. Written by Hui Jiang from the Wong lab at Stanford. Builds available for most OS's.
- SSAHA - SSAHA (Sequence Search and Alignment by Hashing Algorithm) is a tool for rapidly finding near exact matches in DNA or protein databases using a hash table. Developed at the Sanger Centre by Zemin Ning, Anthony Cox and James Mullikin. C++ for Linux/Alpha.
- SWIFT - The SWIFT suit is a software collection for fast index-based sequence comparison. It contains: SWIFT . fast local alignment search, guaranteeing to find epsilon-matches between two sequences. SWIFT BALSAM . a very fast program to find semiglobal non-gapped alignments based on k-mer seeds. Authors are Kim Rasmussen (SWIFT) and Wolfgang Gerlach (SWIFT BALSAM)
- SXOligoSeach - SXOligoSearch is a commercial platform offered by the Malaysian based Synamatix. Will align Illumina reads against a range of Refseq RNA or NCBI genome builds for a number of organisms. Web Portal. OS independent.
- Vmatch - A versatile software tool for efficiently solving large scale sequence matching tasks. Vmatch subsumes the software tool REPuter, but is much more general, with a very flexible user interface, and improved space and time requirements. Essentially a large string matching toolbox. POSIX.

Denovo Assembly

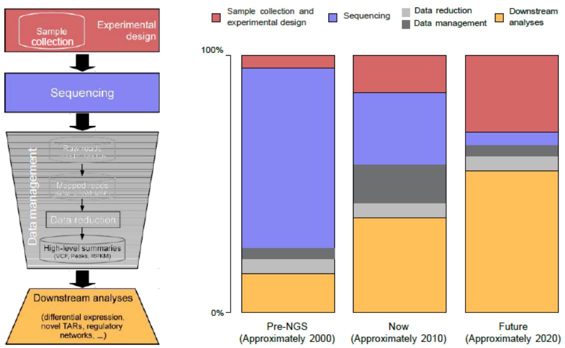
- VELVET - Velvet is a de novo genomic assembler specially designed for short read sequencing technologies, such as Solexa or 454. Need about 20-25X coverage and paired reads. Developed by Daniel Zerbino and Ewan Birney at the European Bioinformatics Institute (EMBL-EBI).
- SOAPdenovo- denovo assenmbler part of SOAP (Short Oligonucleotide Alignment Program).
- MIRA- MIRA (Mimicking Intelligent Read Assembly) is able to perform true hybrid de-novo assemblies using reads gathered through 454 sequencing technology (GS20 or GS FLX). Compatible with 454, Solexa and Sanger data. Linux OS required.
- EULER - Short read de novo assembly. By Mark J. Chaisson and Pavel A. Pevzner from UCSD (published in Genome Research). Uses a de Bruijn graph approach.
- VCAKE - De novo assembly of short reads with robust error correction. An improvement on early versions of SSAKE.
- AMOS - Manipulation of input and output files related to whole-genome shotgun assembly.
- ABySS - Assembly By Short Sequences. ABySS is a de novo sequence assembler that is designed for very short reads. The single-processor version is useful for assembling genomes up to 40-50 Mbases in size. The parallel version is implemented using MPI and is capable of assembling larger genomes. By Simpson JT and others at the Canada's Michael Smith Genome Sciences Centre. C++ as source.
- SHARCGS - De novo assembly of short reads. Authors are Dohm JC, Lottaz C, Borodina T and Himmelbauer H. from the Max-Planck-Institute for Molecular Genetics.
- EDENA - Edena (Exact DE Novo Assembler) is an assembler dedicated to process the millions of very short reads produced by the Illumina Genome Analyzer. Edena is based on the traditional overlap layout paradigm. By D. Hernandez, P. Fran
- CELERA - Celera Assembler can reconstruct long sequences of genomic DNA given the fragmentary data produced by whole-genome shotgun sequencing.
- ALLPATHS Broad institute ftp://ftp.broadinstitute.org/pub/crd/ALLPATHS/Release-LG/)
- Cortex_var - Wellcome, EBI, GAC: http://cortexassembler.sourceforge.net/index_cortex_var.html

More Denovo Assembly

- ELVIRA - High throughput assembly of amplicon reads for virus-sized genomes.
- TIGR - The TIGR Assembler is a tool to assemble large shotgun sequencing projects.
- SSAKE - The Short Sequence Assembly by K-mer search and 3' read Extension (SSAKE) is a genomics application for aggressively assembling millions of short nucleotide sequences by progressively searching for perfect 3'-most k-mers using a DNA prefix tree. Authors are Ren
- ARACHNE - ARACHNE is a program for assembling data from whole genome shotgun sequencing experiments. It was designed for long reads from Sanger sequencing technology, and has been used extensively to assemble many genomes, including many that are large and highly repetitive.
- CLC Genomics Workbench - CLC bio supports the latest technological developments. And of course this includes the very interesting developments within Next Generation Sequencing (NGS).
- FORGE - Whole genome assembler that can combine heterogeneous DNA sequencing technologies. Uses MPI to scale across compute clusters.
- RAY - Ray is a parallel software that computes de novo genome assemblies with next-generation sequencing data.
- Ray is written in C++ and can run in parallel on numerous interconnected computers using the message-passing interface (MPI) standard.
- KNIME - KNIME (Konstanz Information Miner) is a user-friendly and comprehensive open-source data integration, processing, analysis, and exploration platform. (KNIME is looking to MPI enable their software mainly at the request of their drug and life sciences customers)
- LaserGene - Comprehensive Software for DNA & Protein Sequence Analysis, Contig Assembly and Sequence Project Management - Now with expanded Next-Generation Sequence Assembly and Analysis Capability
- SeqMan NGen - Software for Next Generation sequence assembly of Illumina, Roche 454 Life Sciences, ABI SOLiD and Helicos Data
- ALLPATHS - ALLPATHS: De novo assembly of whole-genome shotgun microreads. ALLPATHS is a whole genome shotgun assembler that can generate high quality assemblies from short reads. Assemblies are presented in a graph form that retains ambiguities, such as those arising from polymorphism, thereby providing information that has been absent from previous genome assemblies. Broad Institute.
- SAM - SAM is a Whole Genome Assembly (WGA) Management and Visualization Tool. It provides a generic platform for manipulating, analyzing and viewing WGA data, regardless of input type.
- SEQAN - A Consistency-based Consensus Algorithm for De Novo and Reference-guided Sequence Assembly of Short Reads. By Tobias Rausch and others. C++, Linux/Win.
-

Commercial Genomics Software

- Accelrys – “Pipeline Pilot” – comprehensive NGS data analysis-automated workflows
- Biomatters – “Geneious Suite” of DNA sequence analysis SW solutions (Bioinformatics + molecular biology tools in a single package).
- CLC Bio – Full-service bioinformatics sol. Provider
- DNAnexus – Solution built on Amazon Web Services cloud-based storage-and-analysis
- GenomeQuest – Global sol. And service provider of large-scale genomic SW app
- Geospiza (Subsidiary of Perkin Elmer) – Web based enterprise SW sys.



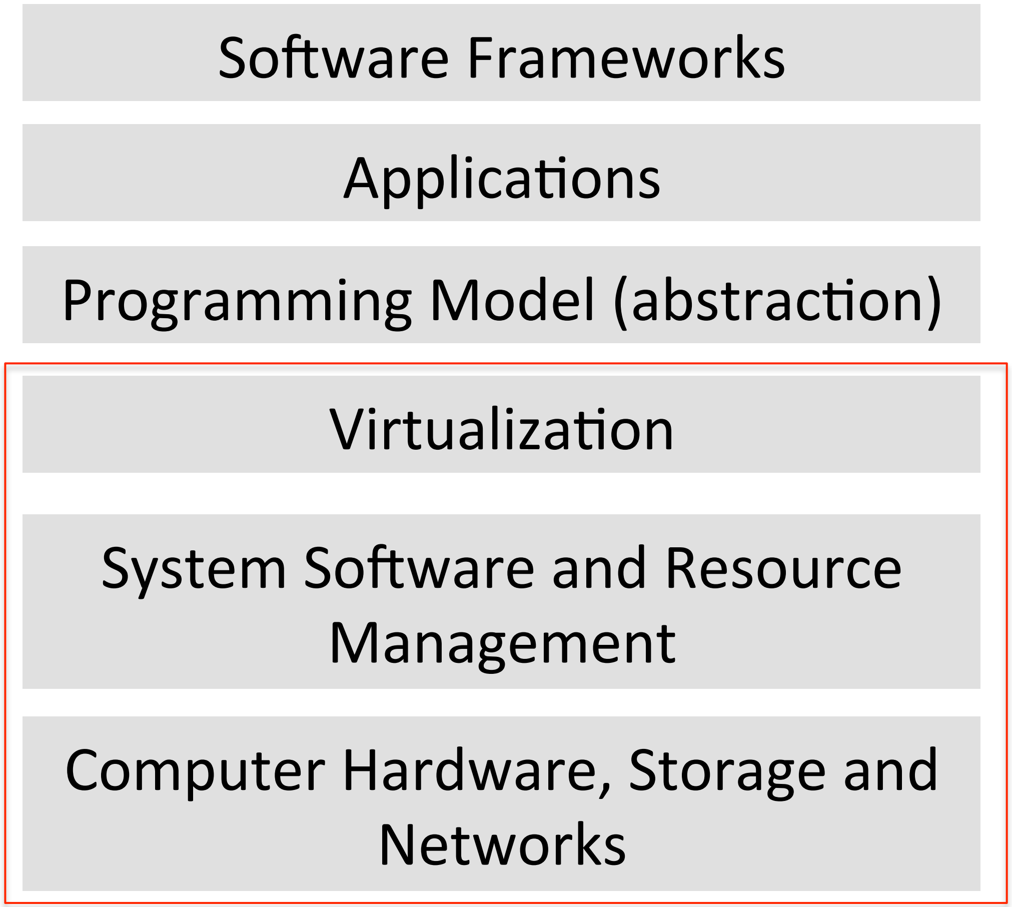
- NextBio – SaaS, Cloud-based scientific platform to aggregate and interpret large quantities of genomic data
- Omixon – NGS Analysis Toolkit; Niche NGS analytic sol. Vendor focused on genomics variant analysis
- Oracle Health Sciences Translational Research Center – Helps clinical researchers normalize, aggregate, and analyze data from variety of sources for “Translational Research” – “Bench to Bedside”
- SAS JMP Genomics – Combines dynamically interactive graphics capabilities of JMP with statistical and analytical power of SAS Analytics.
- Strand Life Sciences – “Avandis” NGS SW solution is a desktop data mining and visualization platform. Focus on small RNA analysis, RNA-Seq transcription analysis, ChIP-Seq transcription regulation analysis. Etc.

Cloud?

- In addition to “tar-balls” for traditional installation on your server (or desktop) all are now freely available as “virtual machines”
  - – that can be loaded into a cloud resource, such as Amazon EC2 (and likely very soon on Google Compute Engine)
  - – Or onto your private cloud implemented by OpenStack, Eucalyptus or VirtualBox.

What is this cloud?

- Compute, storage, networking that you rent – buy it by the pound, as much as you want
  - Amazon EC2, etc., Google Engine
- You do not see the underlying hardware
  - You see a "virtual machine," a software image of a machine that insulates you from the actual hardware
  - Security, portability, resource efficiency
- Package it up and store it away for later retrieval when you don't need it anymore
  - This is called an "instance" in Amazon parlance
- Many organizations layer the top three boxes on top of an instance and repackage it
  - This is what you retrieve and launch when use CloudBioLinux, etc.



Software Frameworks

Applications

Programming Model (abstraction)

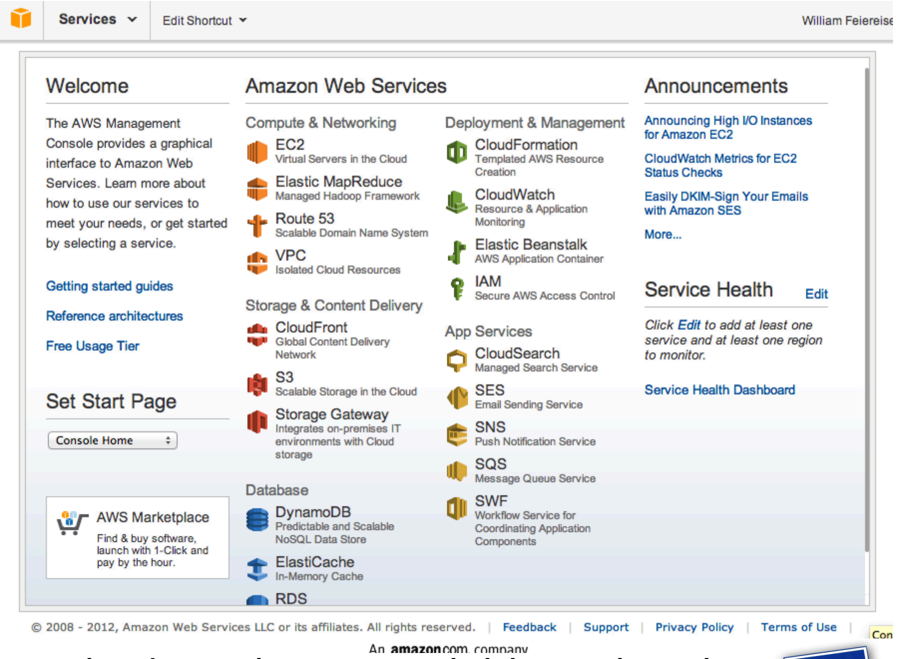
Virtualization

System Software and Resource Management

Computer Hardware, Storage and Networks

Go have a look at Amazon Web Services and the Alestic Images

http://aws.amazon.com/ http://alestic.com/



Services ⌄ | Edit Shortcut ▾ | William Feiereise

Welcome

The AWS Management Console provides a graphical interface to Amazon Web Services. Learn more about how to use our services to meet your needs, or get started by selecting a service.

Getting started guides

Reference architectures

Free Usage Tier

Set Start Page

Console Home ⇕

AWS Marketplace
Find & buy software, launch with 1-Click and pay by the hour.

Amazon Web Services

Compute & Networking

EC2
Virtual Servers in the Cloud

Elastic MapReduce
Managed Hadoop Framework

Route 53
Scalable Domain Name System

VPC
Isolated Cloud Resources

Storage & Content Delivery

CloudFront
Global Content Delivery Network

S3
Scalable Storage in the Cloud

Storage Gateway
Integrates on-premises IT environments with Cloud storage

Database

DynamoDB
Predictable and Scalable NoSQL Data Store

ElastiCache
In-Memory Cache

RDS

Deployment & Management

CloudFormation
Templated AWS Resource Creation

CloudWatch
Resource & Application Monitoring

Elastic Beanstalk
AWS Application Container

IAM
Secure AWS Access Control

App Services

CloudSearch
Managed Search Service

SES
Email Sending Service

SNS
Push Notification Service

SQS
Message Queue Service

SWF
Workflow Service for Coordinating Application Components

Announcements

Announcing High I/O Instances for Amazon EC2

CloudWatch Metrics for EC2 Status Checks

Easily DKIM-Sign Your Emails with Amazon SES

More...

Service Health

Edit

Click Edit to add at least one service and at least one region to monitor.

Service Health Dashboard

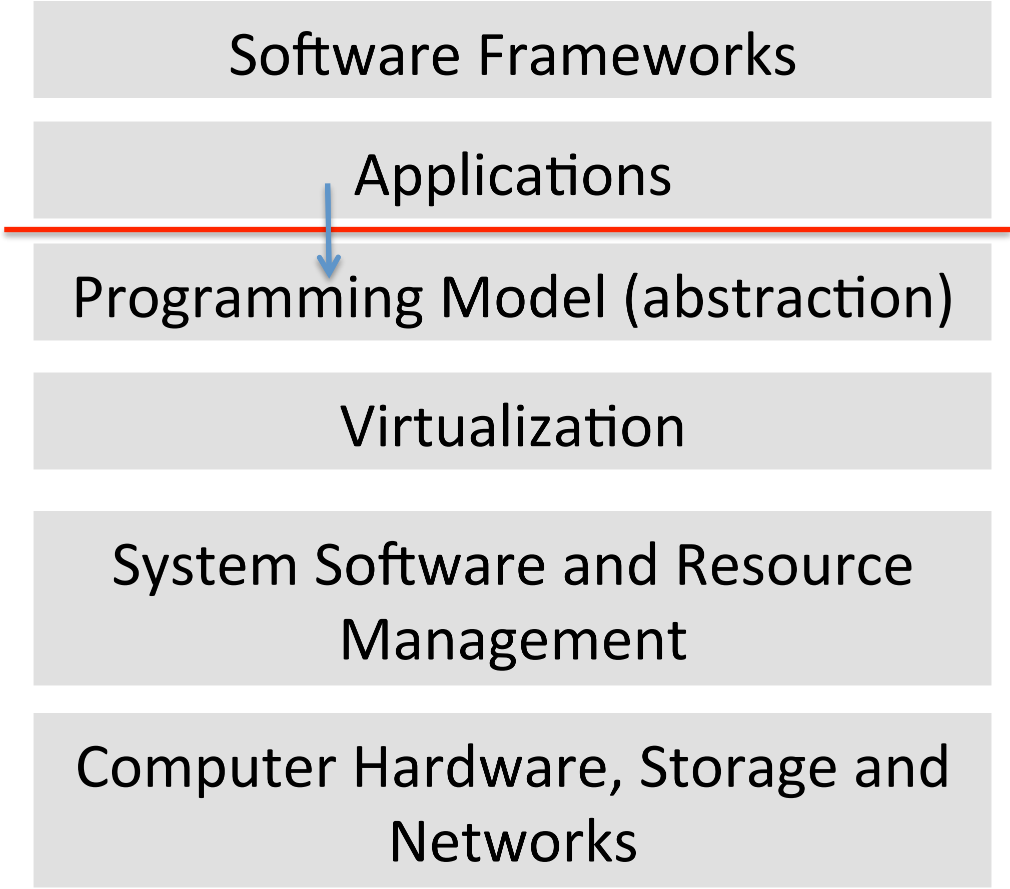
© 2008 - 2012, Amazon Web Services LLC or its affiliates. All rights reserved. | Feedback | Support | Privacy Policy | Terms of Use | Con

An amazon.com. company

- You will be astonished at what is available with only:
- And ***watch Google Compute Engine closely.***
    - They will clearly be integrating their well known services with cloud compute horsepower
    - Their introductory keynote in late June used a human genomics example – calculating in real time

Get some more compute horsepower

- Programming models
  - The view of the underlying compute machinery to which the application is programmed
- For our purposes
  - Simple single processor
    - single threaded – where most bioinformatics apps still are
    - Von Neumann
  - Multiple core
    - Shared memory with multiple threads
    - OpenMP
  - Multiple node
    - Message Passing (MPI)
    - GAS and PGAS
      - (partitioned) global address space
    - An now Map-Reduce (Hadoop) for (unstructured) data



Software Frameworks

Applications

Programming Model (abstraction)

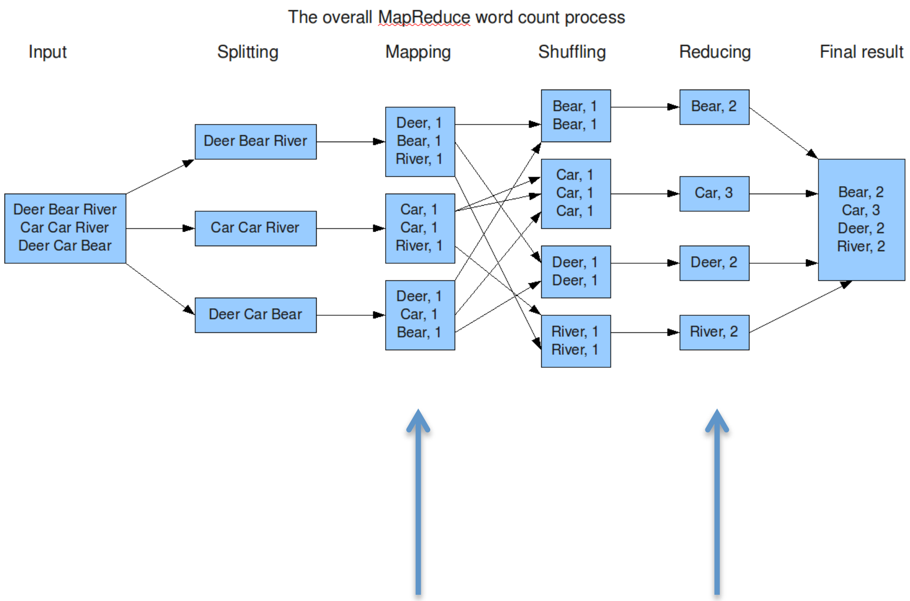
Virtualization

System Software and Resource Management

Computer Hardware, Storage and Networks

Hadoop (aka Map-Reduce)

- For unstructured data Hadoop provides a abstracted and portable way to harness an array of commodity processors in parallel
- For example in a simple word counting exercise →
- Your task is to write two small codes, a mapper and a reducer
- You can download Apache Hadoop and implement it on your favorite hardware (http://hadoop.apache.org/)
- That's fun and useful if run your own data center
- But it's available as Amazon Map-Reduce



The overall MapReduce word count process

Input
Splitting
Mapping
Shuffling
Reducing
Final result

Deer Bear River
Car Car River
Deer Car Bear

Deer Bear River
Car Car River
Deer Car Bear

Deer, 1
Bear, 1
River, 1

Car, 1
Car, 1
River, 1

Deer, 1
Car, 1
Bear, 1

Bear, 1
Bear, 1

Car, 1
Car, 1
Car, 1

Deer, 1
Deer, 1

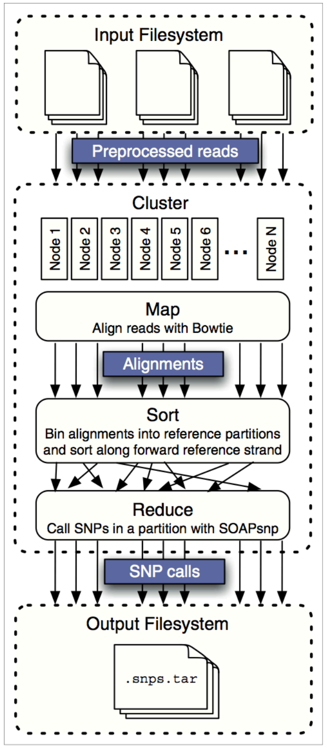
River, 1
River, 1

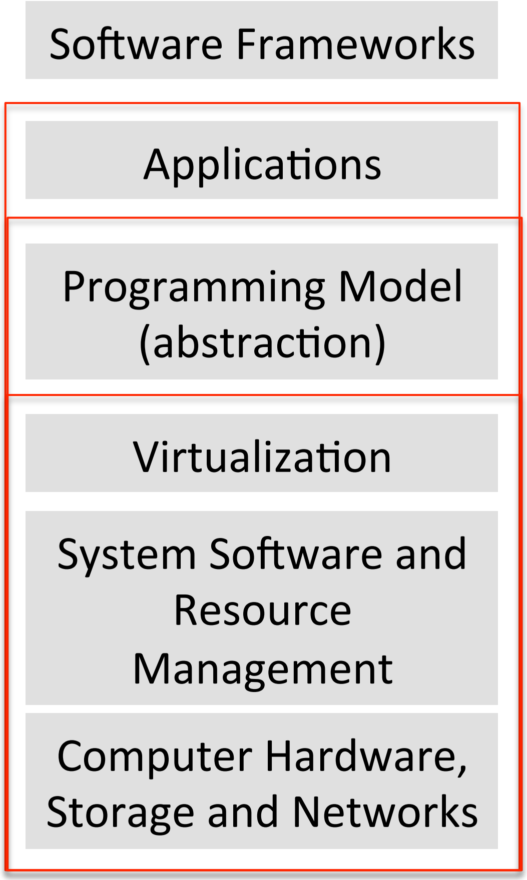
Bear, 2
Car, 3
Deer, 2
River, 2

Bear, 2
Car, 3
Deer, 2
River, 2

Remember Schatz's Cloudburst from earlier?

- That was BLAST on Hadoop
- Schatz and colleagues have now implemented Bowtie and SOAPsnp on Hadoop, called it Crossbow and packaged it as an EC2 cloud image
- Their mapper and reducer are obviously more complicated
- Innovative – you will see more of this





Software Frameworks

Applications

Programming Model (abstraction)

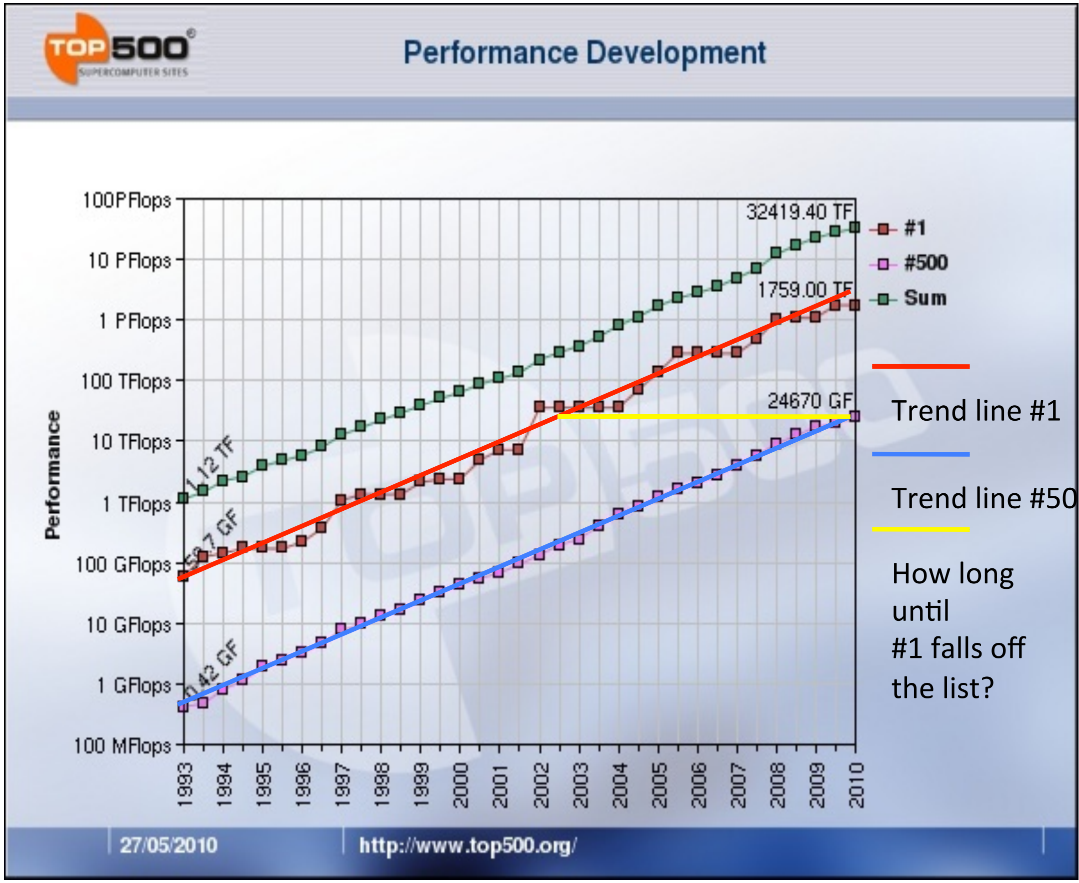
Virtualization

System Software and Resource Management

Computer Hardware, Storage and Networks

And several innovative approaches to the underlying compute machinery -

- But first …
- Much of HPC has concentrated on floating point operations



Trend line #1

Trend line #500

How long until
#1 falls off the list?

Gordon at the San Diego Supercomputing Center

- 1024 compute nodes
  - 64 TB of ***distributed*** memory
- 64 I/O nodes
  - 48 GB of ***distributed*** memory
  - And 256 TB of PCI-attached SSD!
- And there is a 4 PB disk storage system
- Most of the publicized app s/w for Gordon is simulation/ floating point oriented.
- Programming model?
- Remember the SOAPdenovo memory needs from earlier in the talk: 140GB

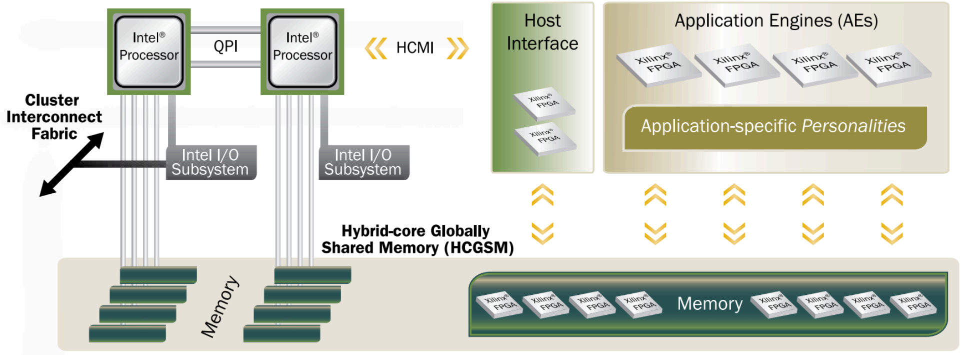
SGI’s Ultra-Violet (UV)

- In its largest shared configuration:
  - 4096 processor cores and 64TB of ***shared memory*** (in 2013)
- Implementations of applications known for large memory requirements
  - Velvet, SOAPdenovo and ABySS
- Excels at implementations of GAS (global address space) as the programming model

Eng Lim Goh (SGI CTO) often points out that the UV is just like a really big PC ☺

Convey Computer

- Field Programmable Gate Arrays (FPGA) augments the general purpose processor
- The programs are called "Personalities" and implement specialized hardware architectures designed to execute specific algorithms.
- Co-processor memory system built around Convey-designed Scatter-Gather DIMMs, optimized for randon transfers of 8-byte bursts, providing near peak bandwidth for non-sequential 8-byte accesses.
- Sharing a logical address space with the general processor main memory
- *I encourage you to speak with the Convey colleagues at this meeting*

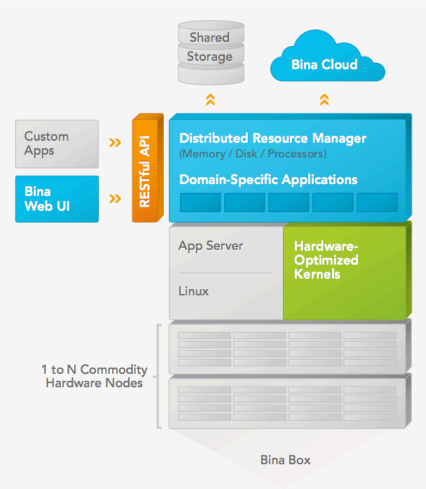

Intel® Processor
QPI
Intel® Processor
HCMI
Host Interface
Application Engines (AEs)
Xilinx® FPGA
Application-specific Personalities
Cluster Interconnect Fabric
Intel I/O Subsystem
Intel I/O Subsystem
Hybrid-core Globally Shared Memory (HCGSM)
Memory
Memory

Standard Intel® x86-64 Server
- x86-64 Linux (RH 6.x)
- Choice of processors, form factor, I/O Chassis, memory size

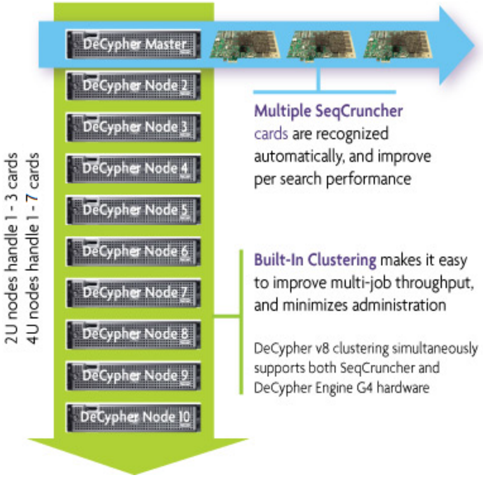
Convey coprocessor
- FPGA-based
- High-bandwidth, highly-parallel physical memory
- Hybrid-core Globally Shared Memory (HCGSM)

Two additional "accelerator" contenders, Bina Box and Timelogic

- Bina Box
- On site computing hardware consisting of graphics processing units, field programmable gate arrays, and multicore central processing unit
- Backed up proprietary backend analysis computing capability and storage in the "Bina Cloud."
- http://www.binatechnologies.com


Shared Storage
Bina Cloud
Custom Apps
Bina Web UI
RESTful API
Distributed Resource Manager
(Memory / Disk / Processors)
Domain-Specific Applications
App Server
Linux
Hardware-Optimized Kernels
1 to N Commodity Hardware Nodes
Bina Box

- Timelogic
- PCIe connected FPGA accelerator
- With implementations of BLAST, SW, Hidden Markov, gene finding s/w



Multiple SeqCruncher cards are recognized automatically, and improve per search performance

Built-In Clustering makes it easy to improve multi-job throughput, and minimizes administration

DeCypher v8 clustering simultaneously supports both SeqCruncher and DeCypher Engine G4 hardware

What we didn't cover – (material for BioHPC 102 ?) Please give me your suggestions

- The instructions issued from compilers that implement the algorithms/applications and their suitability for the underlying processors
  - For example, AVX (in x86), vector instructions for floating point will become AVX2 in the next generation of microarchitecture including integer instructions:
  - And the Fused Multiply-Add (FMA3/4) and XOP instructions
  - MIC, GPGPU and future GP instructions

- Additional fundamental algorithms and their implication for
  - Hidden Markov Models
  - Multiple Sequence Alignment
  - Clustering Tools

- Future memory technologies
  - There's more than just DRAM in our future
  - Connected Solid State Disks
  - Deeper memory hierarchies

Credits

- Materials have been adapted from the following sources:
    - Incogen VIBE E education program
    - NCBI HelpDesk - Field Guide Course Materials
    - Bioinformatics: A practical guide to the analysis of genes and proteins
    - Joanne Fox UBC
    - Mott et. al. from Wellcome Trust
    - George Michaels, Ketan Paranjape – Intel Corp.
    - And much-much material from the following reference list, citations throughout the course material

Thank you

william.j.feiereisen@intel.com or wfeiereisen@cs.unm.edu

Core References

- Afgan, E., Baker, D., Coraor, N., Chapman, B., Nekrutenko, A., & Taylor, J. (2010). Galaxy CloudMan: delivering cloud compute clusters. BMC bioinformatics, 11(Suppl 12), S4. doi: 10.1186/1471-2105-11-S12-S4
- Afgan, E., Chapman, B., Jadan, M., Franke, V., & Taylor, J. (2002). Using Cloud Computing Infrastructure with CloudBioLinux, CloudMan, and Galaxy. (A. D. Baxevanis, G. A. Petsko, L. D. Stein, & G. D. Stormo, Eds.). Hoboken, NJ, USA: John Wiley & Sons, Inc. doi: 10.1002/0471250953.bi1109s38
- Albers, C. A., Lunter, G., MacArthur, D. G., McVean, G., Ouwehand, W. H., & Durbin, R. (2011). Dindel: Accurate indel calls from short-read data. Genome Research, 21(6), 961–973. doi: 10.1101/gr.112326.110
- Arumugam, K., Tan, Y., & Lee, B. (2012). Cloud-enabling Sequence Alignment with Hadoop MapReduce: A Performance Analysis. 2012 4th International Conference on Bioinformatics and Biomedical Technology.
- Baker, B. J., Lesniewski, R. A., & Dick, G. J. (2012). Genome-enabled transcriptomics reveals archaeal populations that drive nitrification in a deep-sea hydrothermal plume. The ISME Journal. doi:10.1038/ismej.2012.64
- Birol, I., Jackman, S. D., Nielsen, C. B., Qian, J. Q., Varhol, R., Stazyk, G., Morin, R. D., et al. (2009). De novo transcriptome assembly with ABySS. Bioinformatics, 25(21), 2872–2877. doi:10.1093/bioinformatics/btp367
- Brüls, T., & Weissenbach, J. (2011). The human metagenome: our other genome? Human Molecular Genetics, 20(R2), R142–R148.
- Butler, J., MacCallum, I., Kleber, M., Shlyakhter, I. A., Belmonte, M. K., Lander, E. S., Nusbaum, C., et al. (2008). ALLPATHS: De novo assembly of whole-genome shotgun microreads. Genome Research, 18(5), 810–820. doi:10.1101/gr.7337908
- Chaisson, M. J., Brinza, D., & Pevzner, P. A. (2008). De novo fragment assembly with short mate-paired reads: Does the read length matter? Genome Research, 19(2), 336–346. doi:10.1101/gr.079053.108
- Chenna, R. (2003). Multiple sequence alignment with the Clustal series of programs. Nucleic acids research, 31(13), 3497–3500. doi:10.1093/nar/gkg500
- Daniel Zerbino Thesis - Genome assembly and comparison using de Bruijn graphs. (n.d.). Daniel Zerbino Thesis - Genome assembly and comparison using de Bruijn graphs.
- Earl, D., Bradnam, K., St John, J., Darling, A., Lin, D., Fass, J., Yu, H. O. K., et al. (2011). Assemblathon 1: a competitive assessment of de novo short read assembly methods. Genome Research, 21(12), 2224–2241. doi:10.1101/gr.126599.111
- Flicek, P., & Birney, E. (2009). Sense from sequence reads: methods for alignment and assembly. Nature Methods, 6(11s), S6–S12. doi:10.1038/nmeth.1376
- Fusaro, V. A., Patil, P., Gafni, E., Wall, D. P., & Tonellato, P. J. (2011). Biomedical Cloud Computing With Amazon Web Services. (F. Lewitter, Ed.)PLoS Comput Biol, 7(8), e1002147. doi:10.1371/journal.pcbi.1002147.t001
- Gertz, J., Varley, K. E., Davis, N. S., Baas, B. J., Goryshin, I. Y., Vaidyanathan, R., Kuersten, S., et al. (2012). Transposase mediated construction of RNA-seq libraries. Genome Research, 22(1), 134–141. doi:10.1101/gr.127373.111
- Goecks, J., Nekrutenko, A., Taylor, J., & Galaxy Team, T. (2010). Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biology, 11(8), R86. doi:10.1186/gb-2010-11-8-r86
- Hillman-Jackson, J., Clements, D., Blankenberg, D., Taylor, J., Nekrutenko, A., & Team, G. (2002). Using Galaxy to Perform Large-Scale Interactive Data Analyses. (A. D. Baxevanis, G. A. Petsko, L. D. Stein, & G. D. Stormo, Eds.).
- Hoboken, NJ, USA: John Wiley & Sons, Inc. doi:10.1002/0471250953.bi1005s38
- Iqbal, Z., Caccamo, M., Turner, I., Flicek, P., & McVean, G. (2012). De novo assembly and genotyping of variants using colored de Bruijn graphs. Nature Genetics, 44(2), 226–232. doi: 10.1038/ng.1028
- Jiménez-Gómez, J. M. (2011). Next generation quantitative genetics in plants. Frontiers in plant science, 2, 77. doi:10.3389/fpls.2011.00077
- Johnson, A. D., Handsaker, R. E., Pulit, S. L., Nizzari, M. M., O'Donnell, C. J., & de Bakker, P. I. W. (2008). SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. Bioinformatics, 24(24), 2938–2939. doi:10.1093/bioinformatics/btn564
- Karakoc, E., Alkan, C., O'Roak, B. J., Dennis, M. Y., Vives, L., Mark, K., Rieder, M. J., et al. (2011). Detection of structural variants and indels within exome data. Nature Methods, 9(2), 176–178. doi:10.1038/nmeth.1810
- Klus, P., Lam, S., Lyberg, D., Cheung, M. S., Pullan, G., McFarlane, I., Yeo, G. S., et al. (2012). BarraCUDA - a fast short read sequence aligner using graphics processing units. BMC research notes, 5, 27. doi:10.1186/1756-0500-5-27
- Krampis, K., Booth, T., Chapman, B., Tiwari, B., Bicak, M., Field, D., & Nelson, K. E. (2012). Cloud BioLinux: pre-configured and on-demand bioinformatics computing for the genomics community. BMC bioinformatics, 13, 42. doi:10.1186/1471-2105-13-42
- Lam, T. W., Sung, W. K., Tam, S. L., Wong, C. K., & Yiu, S. M. (2008). Compressed indexing and local alignment of DNA. Bioinformatics, 24(6), 791–797. doi:10.1093/bioinformatics/btn032
- Langmead, B., Schatz, M. C., Lin, J., Pop, M., & Salzberg, S. L. (2009a). Searching for SNPs with cloud computing. Genome Biology, 10(11), R134. doi:10.1186/gb-2009-10-11-r134
- Langmead, B., Trapnell, C., Pop, M., & Salzberg, S. L. (2009b). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biology, 10(3), R25. doi: 10.1186/gb-2009-10-3-r25
- Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A., McWilliam, H., Valentin, F., et al. (2007). Clustal W and Clustal X version 2.0. Bioinformatics, 23(21), 2947–2948. doi:10.1093/bioinformatics/btm404
- Lesniewski, R. A., Jain, S., Anantharaman, K., Schloss, P. D., & Dick, G. J. (2012). The metatranscriptome of a deep-sea hydrothermal plume is dominated by water column methanotrophs and lithotrophs. The ISME Journal. doi:10.1038/ismej.2012.63
- Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics, 25(14), 1754–1760. doi:10.1093/bioinformatics/btp324
- Li, H., Ruan, J., & Durbin, R. (2008a). Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Research, 18(11), 1851–1858. doi:10.1101/gr.078212.108
- Li, R., Li, Y., Kristiansen, K., & Wang, J. (2008b). SOAP: short oligonucleotide alignment program. Bioinformatics, 24(5), 713–714. doi:10.1093/bioinformatics/btn025
- Li, R., Yu, C., Li, Y., Lam, T. W., Yiu, S. M., Kristiansen, K., & Wang, J. (2009). SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics, 25(15), 1966–1967. doi:10.1093/bioinformatics/btp336
- Li, R., Zhu, H., Ruan, J., Qian, W., Fang, X., Shi, Z., Li, Y., et al. (2010). De novo assembly of human genomes with massively parallel short read sequencing. Genome Research, 20(2), 265–272. doi: 10.1101/gr.097261.109
- Li, W., Fu, L., Niu, B., Wu, S., & Wooley, J. (2012). Ultrafast clustering algorithms for metagenomic sequence analysis. Briefings in bioinformatics. doi:10.1093/bib/bbs035
- Liu, C. M., Wong, T., Wu, E., Luo, R., Yiu, S. M., Li, Y., Wang, B., et al. (2012). SOAP3: ultra-fast GPU-based parallel alignment tool for short reads. Bioinformatics, 28(6), 878–879. doi:10.1093/bioinformatics/bts061

Core References

- Lunter, G., & Goodson, M. (2011). Stampy: A statistical algorithm for sensitive and fast mapping of Illumina sequence reads. Genome Research, 21(6), 936–939. doi:10.1101/gr.111120.110
- Mardis, E. R. (2008). The impact of next-generation sequencing technology on genetics. Trends in genetics, 24(3), 133–141. doi:doi:10.1016/j.tig.2007.12.007
- Markowitz, V. M., Chen, I. M. A., Chu, K., Szeto, E., Palaniappan, K., Grechkin, Y., Ratner, A., Jacob, B., et al. (2011a). IMG/M: the integrated metagenome data management and comparative analysis system. Nucleic acids research, 40(D1), D123–D129. doi:10.1093/nar/gkr975
- Markowitz, V. M., Chen, I. M. A., Palaniappan, K., Chu, K., Szeto, E., Grechkin, Y., Ratner, A., Jacob, B., et al. (2011b). IMG: the integrated microbial genomes database and comparative analysis system. Nucleic acids research, 40(D1), D115–D122. doi:10.1093/nar/gkr1044
- Mende, D. R., Waller, A. S., Sunagawa, S., Järvelin, A. I., Chan, M. M., Arumugam, M., Raes, J., et al. (2012). Assessment of Metagenomic Assembly Using Simulated Next Generation Sequencing Data. (J. Parkinson, Ed.)PLoS ONE, 7(2), e31386. doi:10.1371/journal.pone.0031386.t004
- Miller, J. R., Koren, S., & Sutton, G. (2010a). Assembly algorithms for next-generation sequencing data. Genomics, 95(6), 315–327. doi:10.1016/j.ygeno.2010.03.001
- Miller, J. R., Koren, S., & Sutton, G. (2010b). Assembly algorithms for next-generation sequencing data. Genomics, 95(6), 315–327. doi:10.1016/j.ygeno.2010.03.001
- Opportunistic Data Structures with Applications. (2000). Opportunistic Data Structures with Applications, 1–16.
- Pop, M. (2009a). Genome assembly reborn: recent computational challenges. Briefings in bioinformatics, 10(4), 354–366. doi:10.1093/bib/bbp026
- Pop, M. (2009b). Genome assembly reborn: recent computational challenges. Briefings in bioinformatics, 10(4), 354–366. doi:10.1093/bib/bbp026
- Pop, M. M., & Salzberg, S. L. S. (2008). Bioinformatics challenges of new sequencing technology. Trends in genetics, 24(3), 142–149. doi:10.1016/j.tig.2007.12.006
- Prakash, T., & Taylor, T. D. (2012). Functional assignment of metagenomic data: challenges and applications. Briefings in bioinformatics. doi:10.1093/bib/bbs033
- Riley, D. R., Angiuoli, S. V., Crabtree, J., Dunning Hotopp, J. C., & Tettelin, H. (2012). Using Sybil for interactive comparative genomics of microbes on the web. Bioinformatics, 28(2), 160–166. doi:10.1093/bioinformatics/btr652
- Roberts, A., Trapnell, C., Donaghey, J., Rinn, J. L., & Pachter, L. (2011). Improving RNA-Seq expression estimates by correcting for fragment bias. Genome Biology, 12(3), R22. doi:10.1186/gb-2011-12-3-r22
- Salzberg, S. L., & Ben Langmead. (2012). Fast gapped-read alignment with Bowtie 2. Nature Methods, 1–4. doi:10.1038/nmeth.1923
- Schatz, M. C. (2009). CloudBurst: highly sensitive read mapping with MapReduce. Bioinformatics, 25(11), 1363–1369. doi:10.1093/bioinformatics/btp236
- Schatz, M. C. (n.d.). BlastReduce: high performance short read mapping with MapReduce. University of Maryland, http://cgis. cs. umd. edu/Grad/scholarlypapers/papers/MichaelSchatz. pdf.
- Shi, H., Schmidt, B., Liu, W., & Müller-Wittig, W. (2010). Quality-score guided error correction for short-read sequencing data using CUDA. Procedia Computer Science, 1(1), 1129–1138. doi: 10.1016/j.procs.2010.04.125
- Shimizu, K., & Tsuda, K. (2011). SlideSort: all pairs similarity search for short reads. Bioinformatics, 27(4), 464–470. doi:10.1093/bioinformatics/btq677
- Simpson, J. T., Wong, K., Jackman, S. D., Schein, J. E., Jones, S. J. M., & Birol, I. (2009). ABySS: a parallel assembler for short read sequence data. Genome Research, 19(6), 1117–1123. doi: 10.1101/gr.089532.108
- Special Issue PapersThe many faces of sequence alignment. (2005). Special Issue PapersThe many faces of sequence alignment, 1–17.
- Stein, L. D. (2010). The case for cloud computing in genome informatics. Genome Biology, 11(5), 207. doi:10.1186/gb-2010-11-5-207
- Trapnell, C., & Salzberg, S. L. (2009). How to map billions of short reads onto genomes. Nature biotechnology, 27(5), 455–457.
- Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., Salzberg, S. L., et al. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nature biotechnology, 28(5), 511–515. doi: 10.1038/nbt.1621
- Ye, K., Schulz, M. H., Long, Q., Apweiler, R., & Ning, Z. (2009). Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics, 25(21), 2865–2871. doi:10.1093/bioinformatics/btp394
- Zerbino, D. R., & Birney, E. (2008). Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Research, 18(5), 821–829. doi:10.1101/gr.074492.107
- Zhang, J., Chiodini, R., Badr, A., & Zhang, G. (2011). The impact of next-generation sequencing on genomics. Journal of genetics and genomics = Yi chuan xue bao, 38(3), 95–109. doi:10.1016/j.jgg.2011.02.003

GPGPU References – Systems Biology

- Alhadi Bustamam, Kevin Burrage, Nicholas A. Hamilton (2011) **Fast Parallel Markov Clustering in Bioinformatics using Massively Parallel Computing on GPU with CUDA and ELLPACK-R Sparse Format**. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 99 (PrePrints) IEEE Computer Society. Los Alamitos, CA, USA. [doi]
- Ling Sing Yung, Can Yang, Xiang Wan, Weichuan Yu (2011) **GBOOST: a GPU-based tool for detecting gene–gene interactions in genome–wide case control studies**. *Bioinformatics* 27 (9) 1309-1310. [doi | web]
- Guido Klingbeil, Radek Erban, Mike Giles, Philip Maini (2011) **STOCHSIMGPU: Parallel stochastic simulation for the Systems Biology Toolbox 2 for MATLAB**. *Bioinformatics* 27 (8) [web]
- Yanxiang Zhou, Juliane Liepe, Xia Sheng, Michael P.H. Stumpf, Chris Barnes (2011) **GPU accelerated biochemical network simulation**. *Bioinformatics* [doi | web]
- Juliane Liepe, Chris Barnes, Erika Cule, Kamil Erguler, Paul Kirk, Tina Toni, Michael P.H. Stumpf (2010) **ABC-SysBio—approximate Bayesian computation in Python with GPU support**. 1797-1799. [doi | web]
- M. Vigelius, A. Lane, B. Meyer (2010) **Accelerating Reaction-Diffusion Simulations with General-Purpose Graphics Processing Units**. *Bioinformatics* [doi]

GPGPU References - Sequence comparison

- Chi-Man Liu, Thomas Wong, Edward Wu, Ruibang Luo, Siu-Ming Yiu, Yingrui Li, B. Wang, C. Yu, X. Chu, K. Zhao, Ruiqiang Li, Tak-Wah Lam (2012) **SOAP3: Ultra-fast GPU-based parallel alignment tool for short reads**. *Bioinformatics* [doi | web]
- Panagiotis D. Vouzis, Nikolaos V. Sahinidis (2011) **GPU-BLAST: using graphics processors to accelerate protein sequence alignment**. *Bioinformatics* 27 (2) 182-188. [doi | web]
- Weiguo Liu, Bertil Schmidt, Wolfgang Muller-Wittig (2011) **CUDA-BLASTP: Accelerating BLASTP on CUDA-Enabled Graphics Hardware**. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 8 1678-1684. IEEE Computer Society. Los Alamitos, CA, USA. [doi | web]
- Yongchao Liu, Bertil Schmidt, Douglas Maskell (2010) **CUDASW+2.0: enhanced Smith-Waterman protein database search on CUDA-enabled GPUs based on SIMT and virtualized SIMD abstractions**. *BMC Research Notes* 3 (1) 93. [doi | web]
- Haixiang Shi, B. Schmidt, Weiguo Liu, W. Muller-Wittig (may 2009) **Accelerating error correction in high-throughput short-read DNA sequencing data with CUDA**. In *Parallel Distributed Processing, 2009. IPDPS 2009. IEEE International Symposium on*. 1 -8. [doi]
- Cole Trapnell, Michael C. Schatz (2009) **Optimizing data intensive GPGPU computations for DNA sequence alignment**. *Parallel Computing* 35 (8-9) 429 - 440. [doi | web]
- Yongchao Liu, Bertil Schmidt, Douglas Maskell (2009) **Parallel Reconstruction of Neighbor-Joining Trees for Large Multiple Sequence Alignments using CUDA**. In *IEEE International Workshop on High Performance Computational Biology (HiCOMB 2009)*. 1–8. IEEE Computer Society. Washington, DC, USA. [doi | web]
- Lukasz Ligowski, Witold Rudnicki (2009) **An efficient implementation of Smith Waterman algorithm on GPU using CUDA, for massively parallel scanning of sequence databases**. In *IEEE International Workshop on High Performance Computational Biology (HiCOMB 2009)*. 1-8. IEEE Computer Society. [doi]
- Yongchao Liu, Douglas Maskell, Bertil Schmidt (2009) **CUDASW++: optimizing Smith-Waterman sequence database searches for CUDA-enabled graphics processing units**. *BMC Research Notes* 2 (1) 73. [doi | web]
- Svetlin Manavski, Giorgio Valle (2008) **CUDA compatible GPU cards as efficient hardware accelerators for Smith-Waterman sequence alignment**. *BMC Bioinformatics* 9 (Suppl 2) S10. [doi | web]
- Michael Schatz, Cole Trapnell, Arthur Delcher, Amitabh Varshney (2007) **High-throughput sequence alignment using Graphics Processing Units**. *BMC Bioinformatics* 8 (1) 474. [doi | web]
- Weiguo Liu, Bertil Schmidt, Gerrit Voss, Wolfgang M\"uller-Wittig (2006) **GPU-ClustalW: Using Graphics Hardware to Accelerate Multiple Sequence Alignment**. In *IEEE International Conference on High Performance Computing (HiPC 2006)*. 363–374. Springer Berlin Heidelberg. Berlin, Heidelberg. [doi | web]

GPGPU References - RNA

Guillaume Rizk, Dominique Lavenier (2009) **GPU Accelerated RNA Folding Algorithm**. In *Computational Science – ICCS 2009*. 1004-1013. Springer Berlin / Heidelberg. [doi]

Jens Reeder, Peter Steffen, Robert Giegerich (2007) **pknotsRG: RNA pseudoknot folding including near-optimal structures and sliding windows**. *Nucleic Acids Research* 35 (suppl 2) W320-W324. [doi | web]

David H. Mathews, Jeffrey Sabina, Michael Zuker, Douglas H. Turner (1999) **Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure**. *Journal of Molecular Biology* 288 (5) 911 - 940. [doi | web]

GPGPU References - Phylogeny

- Raphael Helaers, Michel Milinkovitch (2010) **MetaPIGA v2.0: maximum likelihood large phylogeny estimation using the metapopulation genetic algorithm and other stochastic heuristics**. *BMC Bioinformatics* 11 (1) 379. [doi | web]
- Marc A. Suchard, Andrew Rambaut (2009) **Many-core algorithms for statistical phylogenetics**. *Bioinformatics* [doi]
- Maria Charalambous, Pedro Trancoso, Alexandros Stamatakis (2005) **Initial Experiences Porting a Bioinformatics Application to a Graphics Processor**. In *Advances in Informatics*. 415-425. Springer Berlin / Heidelberg. [doi]

GPGPU References - Other

- Ivo D Shterev, Sin-Ho Jung, Stephen L George, Kouros Owzar (2010) **permGPU: Using graphics processing units in RNA microarray association studies**. *BMC Bioinformatics 2010* [doi]
- J.P. Walters, V. Balu, S. Kompalli, V. Chaudhary (may 2009) **Evaluating the use of GPUs in liver image segmentation and HMMER database searches**. In *IEEE International Symposium on Parallel & Distributed Processing (IPDPS'09)*. 1 -12. [doi]
- Elijah Roberts, John Stone, Leonardo Sepulveda, Wen-Mei Hwu, Zaida Luthey-Schulten (2009) **Long Time-scale Simulations of in vivo Diffusion using GPU Hardware**. In *IEEE International Workshop on High Performance Computational Biology (HiCOMB 2009)*. 1–8. IEEE Computer Society. Washington, DC, USA. [doi | web]
- M. Giraud, J.-S. Varre (30 2009-july 4 2009) **Parallel Position Weight Matrices Algorithms**. In *Parallel and Distributed Computing, 2009. ISPDC '09. Eighth International Symposium on*. 65 -72. [doi]
- Nicholas A. Davis, Ahwan Pandey, B. A. McKinney (2011) **Real-world comparison of CPU and GPU implementations of SNPrank: a network analysis tool for GWAS**. 284-285. [doi | web]
- Joshua Buckner, Justin Wilson, Mark Seligman, Brian Athey, Stanley Watson, Fan Meng (2010) **The gputools package enables GPU computing in R**. 134-135. [doi | web]

GPGPU References -Proteomics

- Alex D Stivala, Peter J Stuckey, Anthony I Wirth (2010) **Fast and accurate protein substructure searching with simulated annealing and GPUs,**. *BMC Bioinformatics 2010* [doi]

- D.W. Ritchie, V. Venkatraman (2010) **Ultra-Fast FFT Protein Docking On Graphics Processors.**. *Bioinformatics* [doi]

- Rene Hussong, Barbara Gregorius, Andreas Tholey, Andreas Hildebrandt (2009) **Highly accelerated feature detection in proteomics data sets using modern graphics processing units**. *Bioinformatics* [doi]