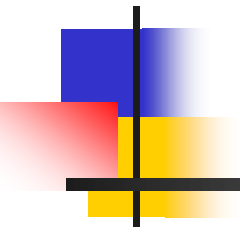
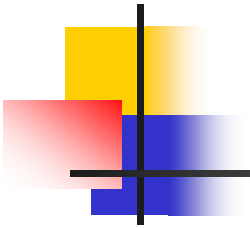


Searching and Sorting

Chen-Hanson Ting

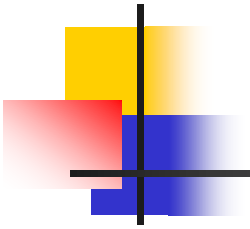
SVFIG

August 25, 2018



Summary

- Genome data
- Optimized Compare
- Optimized Look
- Binary Search
- Optimized Search
- Examples



Genome Data Files

- Gene Bank format
  - plain text
  - Field separators (\, =)
- Fsa, fna formats
  - Annotation starts with >
  - Lines of DNA data
- Working files are all tab delimited text files suitable for Excel

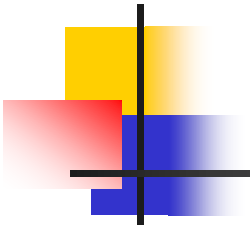
Gene Bank format

```
gene            complement(13363..13743)
                /gene="TDA8"
                /locus_tag="YAL064C-A"
                /gene_synonym="YAL065C-A"
mRNA            complement(13363..13743)
                /gene="TDA8"
                /locus_tag="YAL064C-A"
                /gene_synonym="YAL065C-A"
                /product="Tda8p"
CDS             complement(13363..13743)
                /gene="TDA8"
                /locus_tag="YAL064C-A"
                /gene_synonym="YAL065C-A"
                /GO_component="GO:0005575 - cellular_component [Evidence
                ND]"
                /GO_function="GO:0003674 - molecular_function [Evidence
                ND]"
                /GO_process="GO:0008150 - biological_process [Evidence
                ND]"
                /note="hypothetical protein; null mutant is sensitive to
                expression of the top1-T722A allele; not an essential
                gene"
                /codon_start=1
                /product="Tda8p"
                /protein_id="DAA06923.1"
                /db_xref="SGD:S000002140"
                /translation="MTGYFLPPQTSSYTFRFAKVDDSAILSVGGDVAFGCCAQEQPPI
                TSTNFTINGIKPWQGRLPDNIAGTVYMYAGFYCPMKIVYSNAVSWHTLPVSVELPDVT
                TVSDDFAGHVYSFDDDLTAQLYYP"
gene            21566..21850
```

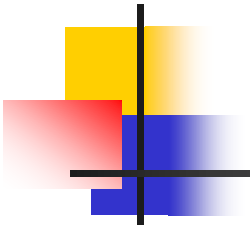
Gene Bank format

```
BASE COUNT    69836 a  44641 c  45766 g  69975 t
ORIGIN
        1 ccacaccaca cccacacacc cacacaccac accacacacc acaccacacc cacacacaca
       61 catcctaaca ctaccctaac acagccctaa tctaaccctg gccaacctgt ctctcaactt
      121 accctccatt accctgcctc cactcgttac cctgtcccat tcaaccatac cactccgaac
      181 caccatccat ccctctactt actaccactc acccaccgtt accctccaat tacccatatc
      241 caacccactg ccacttaccc taccattacc ctaccatcca ccatgaccta ctcaccatac
      301 tgttcttcta cccaccatat tgaaacgcta acaaatgatc gtaaataaca cacacgtgct
      361 taccctacca ctttatacca ccaccacatg ccatactcac cctcacttgt atactgattt
      421 tacgtacgca cacggatgct acagtatata ccatctcaaa cttaccctac tctcagattc
      481 cacttcactc catggcccat ctctcactga atcagtacca aatgcactca catcattatg
      541 cacggcactt gcctcagcgg tctataccct gtgccattta cccataacgc ccatcattat
      601 ccacattttg atatctatat ctcattcggc ggtcccaaat attgtataac tgcccttaat
      661 acatacgtta taccactttt gcaccatata cttaccactc catttatata cacttatgtc
      721 aatattacag aaaaatcccc acaaaaatca cctaaacata aaaatattct acttttcaac
      781 aataatacat aaacatattg gcttgtggta gcaacactat catggtatca ctaacgtaaa
      841 agttcctcaa tattgcaatt tgcttgaacg gatgctattt cagaatattt cgtacttaca
      901 caggccatac attagaataa tatgtcacat cactgtcgta acactcttta ttcaccgagc
      961 aataatacgg tagtggctca aactcatgcg ggtgctatga tacaattata tcttatttcc
     1021 attcccatat gctaaccgca atatcctaaa agcataactg atgcatcttt aatcttgtat
     1081 gtgacactac tcatacgaag ggactatatc tagtcaagac gatactgtga taggtacgtt
     1141 atttaatagg atctataacg aaatgtcaaa taattttacg gtaatataac ttatcagcgg
     1201 cgtatactaa aacggacgtt acgatattgt ctcacttcat cttaccaccc tctatcttat
     1261 tgctgataga acactaaccc ctcagcttta tttctagtta cagttacaca aaaaactatg
     1321 ccaacccaga aatcttgata ttttacgtgt caaaaaatga gggtctctaa atgagagttt
     1381 ggtaccatga cttgtaactc gcactgccct gatctgcaat cttgttctta gaagtgacgc
     1441 atattctata cggcccgacg cgacgcgcca aaaaatgaaa aacgaagcag cgactcattt
```

Fsa Format



```
>tpg|BK006935.2| [organism=Saccharomyces cerevisiae
S288c] [strain=S288c] [moltype=genomic] [chromosome=I]
[note=R64-1-
1]CCACACCACACCCACACACCCACACACCACACCACACACCACACCACACCCAC
ACACACACATCCTAACACTACCCTAACACAGCCCTAATCTAACCCTGGCCAACCT
GTCTCTCAACTTACCCTCCATTACCCTGCCTCCACTCGTTACCCTGTCCCATTCA
ACCATACCACTCCGAACCACCATCCATCCCTCTACTTACTACCACTCACCCACCG
TTACCCTCCAATTACCCATATCCAACCCACTGCCACTTACCCTACCATTACCCTA
CCATCCACCATGACCTACTCACCATACTGTTCTTCTACCCACCATATTGAAACGC
TAACAAATGATCGTAAATAACACACACGTGCTTACCCTACCACTTTATACCACCA
CCACATGCCATACTCACCCTCACTTGTATACTGATTTTACGTACGCACACGGATG
CTACAGTATATACCATCTCAAACTTACCCTACTCTCAGATTCCACTTCACTCCAT
GGCCCATCTCTCACTGAATCAGTACCAAATGCACTCACATCATTATGCACGGCAC
TTGCCTCAGCGGTCTATACCCTGTGCCATTTACCCATAACGCCCATCATTATCCA
CATTTTGATATCTATATCTCATTCGGCGGTCCCAAATATTGTATAACTGCCCTTA
```



Optimize Searching

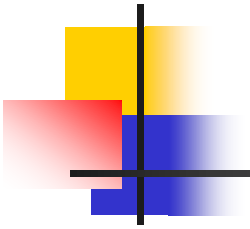
- **Intel 80x86 is a native searching machine.**
- **It has this very powerful instruction pair REPZ CMPSB, which does string comparison automatically.**
- **SIMD? Multiple Cores?**

Optimized Compare

```
code CompareStr ( string pattern n -- -1 | 0 | 1 )
( 044730 8B CB             )   mov     ecx , ebx
( 044732 8B 7D 00          )   mov     edi , 0 [ebp]
( 044735 8B 75 04          )   mov     esi , 4 [ebp]
( 044738 8D 6D 08          )   lea     ebp , 8 [ebp]
( 04473B 33 DB             )   xor     ebx , ebx
( 04473D 81 E1 FF 00 00 00 )   and     ecx , # dword $FF
( 044743 F3                )   repz
( 044744 A6                )   cmpsb
( 044745 77 03             )   ja      short @@1
( 044747 72 04             )   jb      short @@2
( 044749 C3                )   ret     near
( 04474A                   ) @@1:
( 04474A FF C3             )   inc     ebx
( 04474C C3                )   ret     near
( 04474D                   ) @@2:
( 04474D FF CB             )   dec     ebx
( 04474F C3                )   ret     near
end-code
```

Optimized Look

```
code look ( pattern n1 string n2 -- match|0 )
( 65651E4                    ) @@1:
( 65651E4 FF CB              )   dec     ebx
( 65651E6 7E 19              )   jle     short @@3
( 65651E8 8B 7D 00           )   mov     edi , 0 [ebp]
( 65651EB 8B 4D 04           )   mov     ecx , 4 [ebp]
( 65651EE 8B 75 08           )   mov     esi , 8 [ebp]
( 65651F1 F3                 )   repz
( 65651F2 A7                 )   cmpsd
( 65651F3 75 07              )   jnz     short @@2
( 65651F5 8B 5D 00           )   mov     ebx , 0 [ebp]
( 65651F8 8D 6D 0C           )   lea     ebp , $C [ebp]
( 65651FB C3                 )   ret     near
( 65651FC                    ) @@2:
( 65651FC FF 45 00           )   inc     0 [ebp] dword
( 65651FF EB E3              )   jmp     short @@1
( 6565201                    ) @@3:
( 6565201 8D 6D 0C           )   lea     ebp , $C [ebp]
( 6565204 C3                 )   ret     near
end-code
```



Optimized Sorting

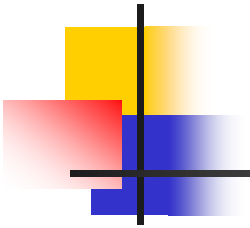
- **Excel has excellent sorting capabilities.**
- **Convert data files to the tab limiting text files, and use Excel to operate on data.**

MicroRNA File

- **miRBase Database**
  - **48,885 miRNA's**
  - **27,790 unique miRNA's**
  - **5312 human miRNA's**

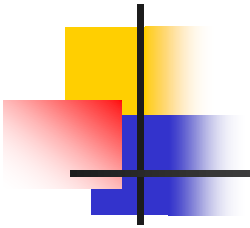
microRNA File

L27299	AAAAAAAAATCATGCTGCATT
L06881	AAAAAAACCTTTTCATGACTTA
L35469	AAAAAAAGATGCAGGACTAGA
L11999	AAAAAAAGGGAAAAGTTTTT
L41666	AAAAAAATGATGGTTCAGG
M17705	AAAAAACAAGGATCCACGGAT
M28136	AAAAAACACCGTTTCCTCCAG
M46624	AAAAAACCGAGTGGACTTTTTG
M18231	AAAAAACTTACGGATCAAGTTGAT
L37526	AAAAAACTTCTACAAAAATAA
M43767	AAAAAACTTTACACCGTCGGT
L25307	AAAAAAGCCACGCTGATACTCT
L28788	AAAAAAGCGGCTTGTACAAAT
L12033	AAAAAAGGAAAACAAACAGCAC
M40980	AAAAAAGGGAATGGCTAAACTTG



Genome Files

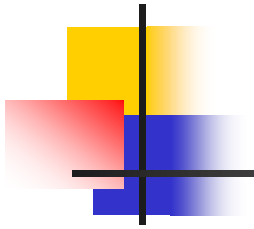
- Stripe off all annotation and formatting characters.
- Data file has only ACGT characters.
- Data read into a map array, and can be read and written at will.



MicroRNA Analysis

- Genome file is scanned to find matching miRNA patterns.
- miRNA records are sorted to facilitate binary search.
- Genome addresses and ID's of matching miRNA are written to output file.

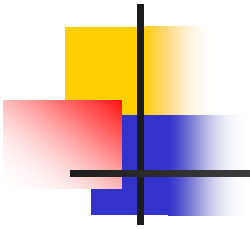
MicroRNA Analysis



- **A list of genome addresses and ID's of matching miRNA is analyzed to identify clusters of consecutive miRNAs.**
- **Clusters of miRNAs might represent complex cell functions.**

microRNA File

11565	YAL065C hypothetical protein	
11685	Y00028	9539
11743	Y00075	58
11951	YAL065C hypothetical protein	
11969	Y00076	226
11999	Y00022	30
12046	YAL064W-B hypothetical protein	
12059	Y00021	60
12082	Y00038	23
12118	Y00070	36
12150	Y00061	32
12190	Y00066	40
12269	Z01320	79
12408	Y00045	139
12426	YAL064W-B hypothetical protein	



Binary Search

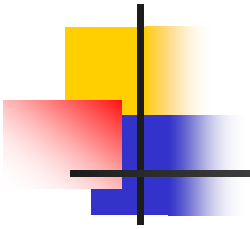
- **miRNA records are variable in length, and binary search does not work directly.**
- **A link table must be created to enable binary search.**

Genomes Studied

Nasuia	112,091
Ruddii	173,806
Equitans	490,885
Genitalium	580,076
E coli	4,641,652
Yeast	12,100,000
Elegans	100,300,000
Mouse	2,700,000,000
Human	3,289,000,000

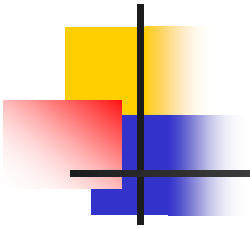
Bacterial Genomes

- **Bacterial genomes do not contain miRNAs in my database.**
- **There are studies on miRNAs in bacteria, but not extensively collected.**



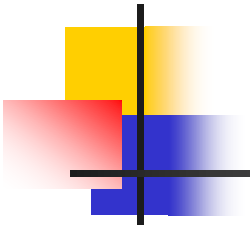
Bacterial Genomes

- **Bacterial genomes are small enough so that I can do exhaustive searches on repeated patterns.**
- **It took a long time, 10 days for e. coli, to search all repeated 20 base patterns.**



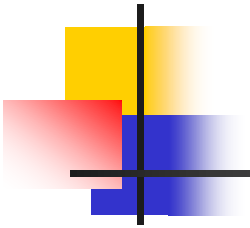
Optimized Search

- For large genomes, I generate 256 lists, each links all patterns with the same 4 starting bases.
- For each 20 base patterns, I only have to search one link, and speed-up searching by 256 times.



Optimized Search

- A link list contain only offsets from one 4 base pattern to the next.
- Offsets are encoded by consecutive 7 bit values, with the MSB as continuation bit.
- 256 link list is slightly larger than genome file.

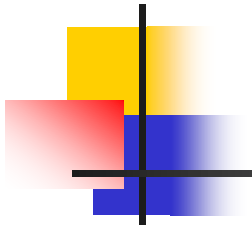


Optimized Search

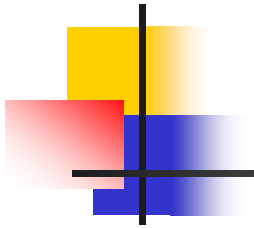
- **Searching time for E. coli (4.6M bp) is reduced from 10 days to 30 minutes.**
- **I think I can now handle much larger genomes.**

Chr01 of Yeast

0	Y01816		11565	YAL065C hypothetica		24000	YAL063C Flo9p		159793	YAR008W Sen34p		181254	tL(CAAA tRNA-Leu				
1804	Y00041	1804	11685	Y00028	9539	24117	Y00028	6043	160237	Y00761	8138	182522	tS(AGAA tRNA-Ser				
1807	YAL068C Pau8p		11743	Y00075	58	24175	Y00075	58	160257	Y02534	20	182576	Y00750	1442			
1855	Y00048	51	11951	YAL065C hypothetica		24401	Y00076	226	160279	Y02408	22	182603	tS(AGAA tRNA-Ser				
1885	Y01509	30	11969	Y00076	226	24431	Y00022	30	160300	Z01533	21	182619	Y02617	43			
2041	Z01877	156	11999	Y00022	30	24491	Y00021	60	160323	Z02433	23	182642	Y01237	23			
2146	Y00047	105	12046	YAL064W-B hypothe		24514	Y00038	23	160348	Y01615	25	182667	Y02510	25			
2169	YAL068C Pau8p		12059	Y00021	60	24550	Y00070	36	160376	Y02319	28	182688	Z00981	21			
2480	YAL067W-A hypothe		12082	Y00038	23	24582	Y00061	32	160401	Y02635	25	182731	Y02231	43			
2707	YAL067W-A hypothe		12118	Y00070	36	24622	Y00066	40	160423	Y02351	22	182755	Y02594	24			
7235	YAL067C Seo1p		12150	Y00061	32	24722	Z01320	100	160469	Y00955	46	182775	Y00482	20			
9016	YAL067C Seo1p		12190	Y00066	40	24861	Y00045	139	160497	Z01159	28	182798	Y02625	23			
11565	YAL065C hypothetica		12269	Z01320	79	24975	Y00046	114	160531	Y01880	34	182818	Y02370	20			
11685	Y00028	9539	12408	Y00045	139	25071	Y00063	96	160554	Z02129	23	182853	Z01367	35			
11743	Y00075	58	12426	YAL064W-B hypothe		25119	Z00071	48	160594	Z02507	40	182874	Y02597	21			
11951	YAL065C hypothetica		12499	Y00005	91	25151	Y00050	32	160597	YAR009C hypothetica		182898	Y01721	24			
11969	Y00076	226	12529	Y00064	30	25476	Y00044	325	160615	Y00682	21	182930	Y01425	32			
11999	Y00022	30	12570	Y00046	41	25530	Z00026	54	160652	Z00596	37	183141	Y02616	211			
12046	YAL064W-B hypothe		12620	Y00015	50	25557	Y00011	27	160680	Z01902	28	183165	Z01901	24			
12059	Y00021	60	12668	Y00015	48	25584	Y00056	27	160701	Y01171	21	183186	Y00722	21			
12082	Y00038	23	12714	Y00046	46	25611	Y00044	27	160744	Y02556	43	183239	Y01536	53			
12118	Y00070	36	12764	Y00015	50	25632	Z0000Z	21	160820	Y00469	76	183285	Z02347	46			
12150	Y00061	32	12787	Y00005	23	25671	Z01333	39	160870	Y02136	50	183306	Y02003	21			
12190	Y00066	40	12821	Y00019	34	25713	Y00036	42	160900	Y02080	30	183394	Z01613	88			
12269	Z01320	79	12906	Y00063	85	25744	Z00033	31	160927	Z01502	27	183419	Z01470	25			



Questions?



Thank You!