Local Alignment and Linear Space Alignment

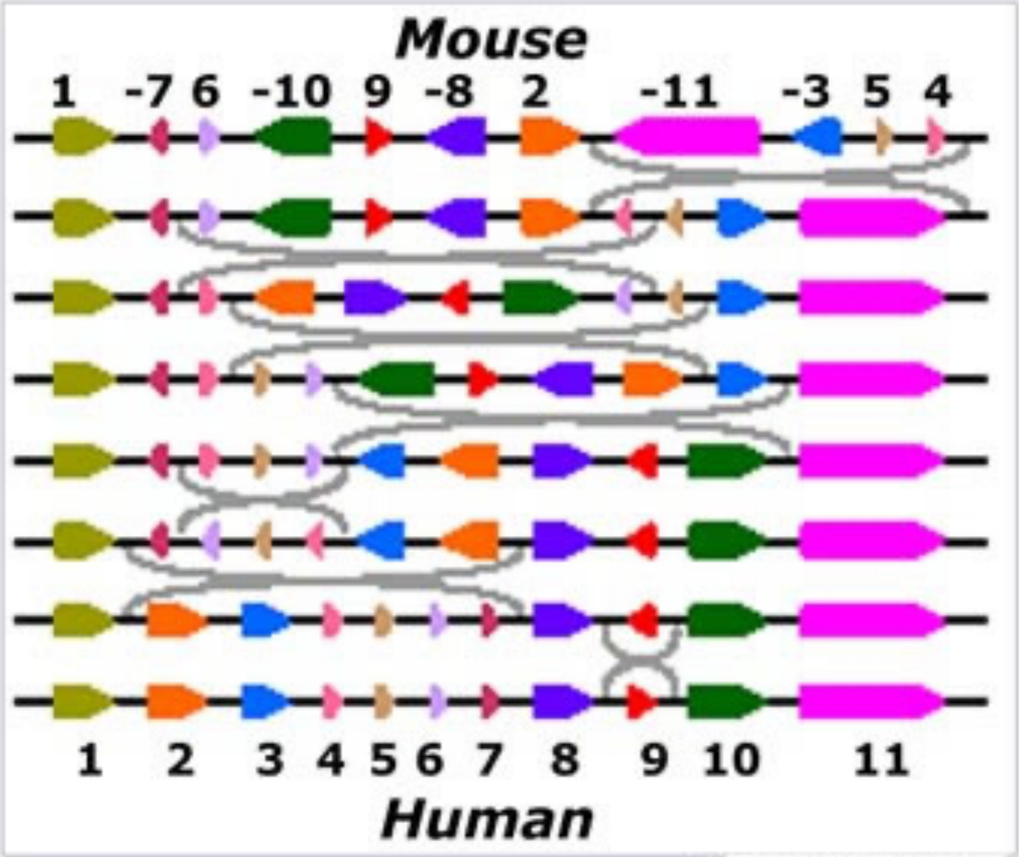
Myoglobin Genes of Mouse and Human

```
>NM_013593.3 Mus musculus myoglobin (Mb), transcript variant 2, mRNA
TCGGGAACTGTTTTAGAAACAGAACATCATCTTCAACATCCAGAGGACTGTCATCCTTGTCCCTGTGGGT
GAGGGAAACAAACACTTGGCTTCAATGTCCCAGGAGAAAGACCCAATTGCTCATCCAGCCCACGTGGCCT
CCAGAAGCCACCATGGGGCTCAGTGATGGGGAGTGGCAGCTGGTGCTGAATGTCTGGGGGAAGGTGGAGG
CCGACCTTGCTGGCCATGGACAGGAAGTCCTCATCGGTCTGTTTAAGACTCACCCTGAGACCCTGGATAA
GTTTGACAAGTTCAAGAACTTGAAGTCAGAGGAAGATATGAAGGGCTCAGAGGACCTGAAGAAGCATGGT
TGCACCGTGCTCACAGCCCTGGGTACCATCCTGAAGAAGAAGGGACAACATGCTGCCGAGATCCAGCCTC
TAGCCCAATCACACGCCACCAAGCACAAGATCCCGGTCAAGTACCTGGAGTTTATCTCAGAAATTATCAT
TGAAGTCCTGAAGAAGAGACATTCCGGGGACTTTGGAGCAGATGCTCAGGGCGCCATGAGCAAGGCCCTG
GAGCTCTTCCGGAATGACATTGCCGCCAAGTACAAGGAGCTAGGCTTCCAGGGCTGAGCCATGGGCTCCC
ACTGTCCAGCCCACCAAGCTGGGACCCAGTGTTGTGTAGCAAGTAGCGTGTGCAGTGTTCTAGGTTAGCA
GAGAACAGAAGAGGGGAGCATAGTGTGGCATCCACCCACACCCCTGGGGACAGGGCTCTGGGCAGTGTTA
CCCTGGAGCCCAGAGGTGCAAAGTGGCCTTCGTCAGCTCTGCCGGGTCATGCTCAGGTCTCCTAAGTCCC
AGTCCATTTTCTTCTGGTTTTGGGAAAATCTCTTTTCCACTGTCACATTTGACCCCAAATCCAAGTCACT
GACTAGCAGACCCTGACCTTTGGGCGAGATGGAGGGTTGCTTAGAGGGAGTGGAGGGTGAAAACGGGGCG
GTGAGCATCAAGTCTCCCACTGCTCAGCTTCCCGTTGACCCACCTTGTCTCAATAAAATATCCTGCGAGT
CCTCAAAAAAAAAAAAAAA
```

```
>NM_005368.3 Homo sapiens myoglobin (MB), transcript variant 1, mRNA
AAACCCCAGCTGTTGGGGCCAGGACACCCAGTGAGCCCATACTTGCTCTTTTTGTCTTCTTCAGACTGCG
CCATGGGGCTCAGCGACGGGGAATGGCAGTTGGTGCTGAACGTCTGGGGGAAGGTGGAGGCTGACATCCC
AGGCCATGGGCAGGAAGTCCTCATCAGGCTCTTTAAGGGTCACCCAGAGACTCTGGAGAAGTTTGACAAG
TTCAAGCACCTGAAGTCAGAGGACGAGATGAAGGCGTCTGAGGACTTAAAGAAGCATGGTGCCACCGTGC
TCACCGCCCTGGGTGGCATCCTTAAGAAGAAGGGGCATCATGAGGCAGAGATTAAGCCCCTGGCACAGTC
GCATGCCACCAAGCACAAGATCCCCGTGAAGTACCTGGAGTTCATCTCGGAATGCATCATCCAGGTTCTG
CAGAGCAAGCATCCCGGGGACTTTGGTGCTGATGCCCAGGGGGCCATGAACAAGGCCCTGGAGCTGTTCC
GGAAGGACATGGCCTCCAACTACAAGGAGCTGGGCTTCCAGGGCTAGGCCCCTGCCGCTCCCACCCCCAC
CCATCTGGGCCCCGGGTTCAAGAGAGAGCGGGGTCTGATCTCGTGTAGCCATATAGAGTTTGCTTCTGAG
TGTCTGCTTTGTTTAGTAGAGGTGGGCAGGAGGAGCTGAGGGGCTGGGGCTGGGGTGTTGAAGTTGGCTT
TGCATGCCCAGCGATGCGCCTCCCTGTGGGATGTCATCACCCTGGGAACCGGGAGTGGCCCTTGGCTCAC
TGTGTTCTGCATGGTTTGGATCTGAATTAATTGTCCTTTCTTCTAAATCCCAACCGAACTTCTTCCAACC
TCCAAACTGGCTGTAACCCCAAATCCAAGCCATTAACTACACCTGACAGTAGCAATTGTCTGATTAATCA
CTGGCCCCTTGAAGACAGCAGAATGTCCCTTTGCAATGAGGAGGAGATCTGGGCTGGGCGGGCCAGCTGG
GGAAGCATTTGACTATCTGGAACTTGTGTGTGCCTCCTCAGGTATGGCAGTGACTCACCTGGTTTTAATA
AAACAACCTGCAACATCTCA
```

Try align them at
EMBL-EBI

Interesting Fact

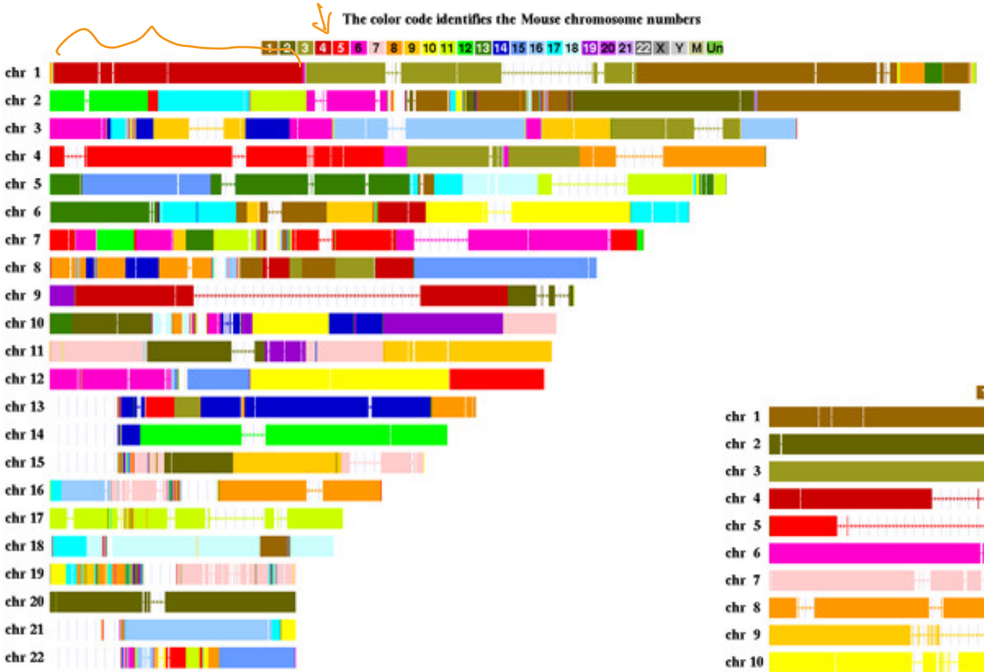


© Genome Research

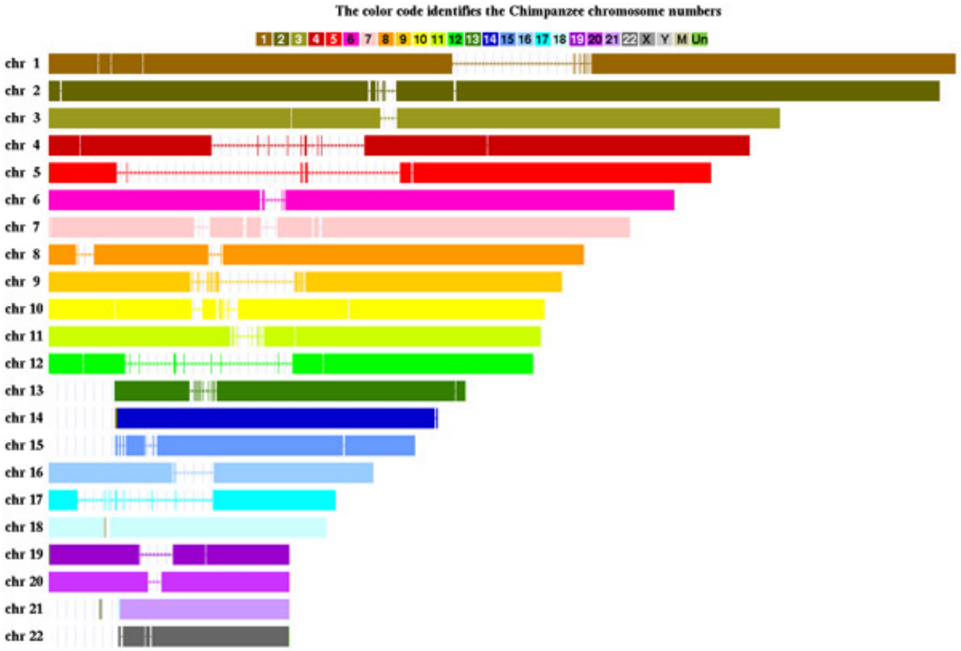
- Human and mouse share big blocks on their genomes.
- Figure shows relation between chromosome X of mouse and human.
- Each colored block is relatively conserved, but different in orders and orientations.
- Seven inversions are required to put them in the correct order and orientation. This is called "sorting by reversals".

pavel pevzner

Mouse, Human, Chimpanzee



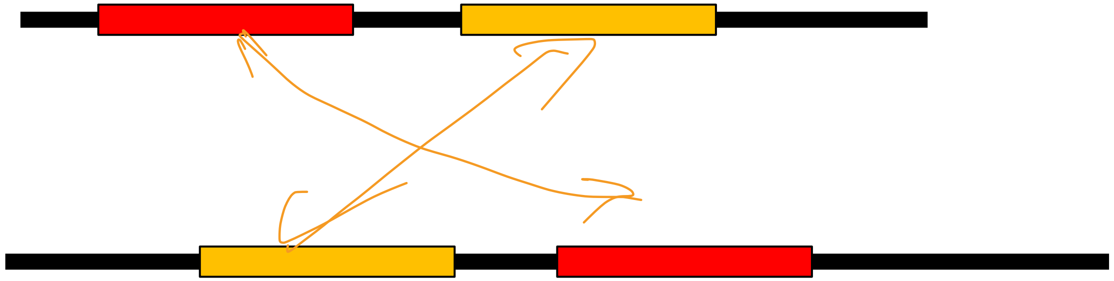
Mouse to Human



Chimpanzee to Human

Local Alignment

- Conserved regions are "local" to the genome/chromosome. But previous alignment is "global".
- We need a model to define "local" similarity.



Local Alignment

- Given: two sequences S and T
- Find: substrings of S and T that maximizes the alignment score.

```
• AATTAG-CCGATGAC
•     || | |||
•  TGGAGGCTGATATA
```

- I.e., The indels at the beginning and end of the two strings are free.

Local Alignment

- Local alignment score is at least 0.
- The model only makes sense for alignment but not edit distance nor LCS.
- Question: Is the optimal local alignment a local part of an optimal “global” alignment?

match = 1   mismatch = -1 ,   indel = -5

A T	A
T A	A
global	local

Warm-up: Prefix alignment

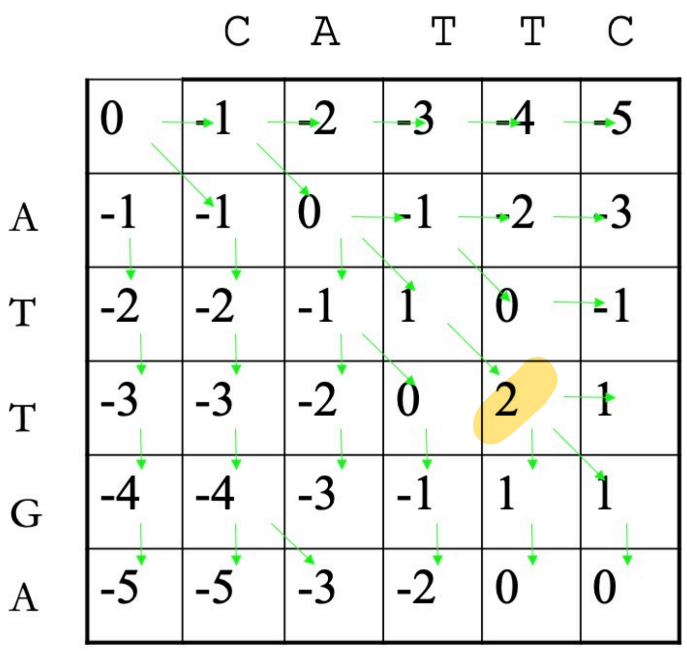
What if we want to find the highest-scoring alignment between two prefixes of the two sequences.

S

T

- CATTC
- ATTGA

Match=1
Mismatch=-1
Indel=-1



		C	A	T	T	C
	0	-1	-2	-3	-4	-5
A	-1	-1	0	-1	-2	-3
T	-2	-2	-1	1	0	-1
T	-3	-3	-2	0	2	1
G	-4	-4	-3	-1	1	1
A	-5	-5	-3	-2	0	0

i, j, to maximize

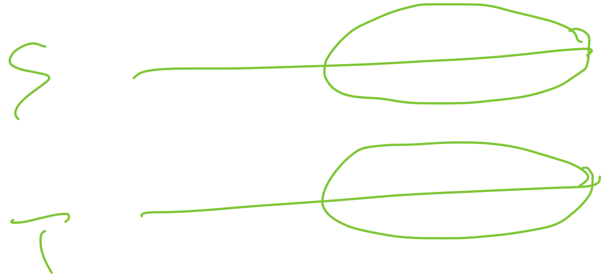
align score(S[1..i], T[1..j])

D[i, j]

$\max_{i,j} D[i,j]$

Warm-up: “suffix alignment”

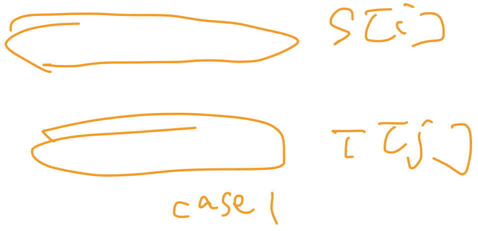
- Suppose we only get the “free” deletions at the prefixes of the alignment.
  - `AATTAG-CCGAT`
  - `    || | |||`
  - ` TGGAGGCTGAT`

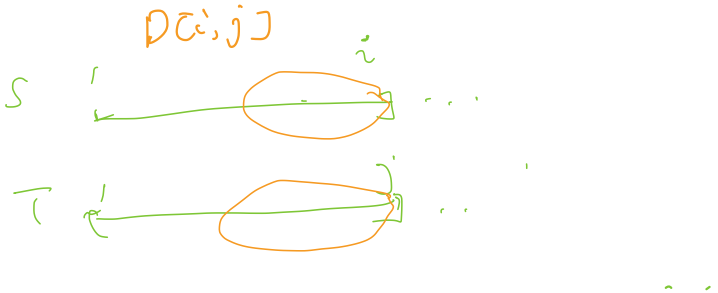


- That is, we choose two suffixes, and align them together optimally.

Last column

- Let D[i,j] denote the optimal “suffix alignment” alignment score of s[1..i], t[1..j].
- That is, D[i,j] is the maximum alignment score for s[i’..i] and t[j’..j] for all i’ and j’.
- Consider the last column of this optimal “suffix” alignment. Four cases arise:





Case 1: s[i] v.s. t[j]

Case 2: s[i] v.s. –

Case 3: t[j] v.s. –

Case 4: an empty alignment

- Case 4 is the only new case comparing to the basic alignment.

DP algorithm for “suffix alignment”

$$D[i,j] = \max \begin{cases} D[i-1, j-1] + f(s[i], t[j]); & \checkmark \\ D[i-1, j] + f(s[i],-); & \checkmark \\ D[i, j-1] + f(-, s[j]); & \checkmark \\ 0 & \checkmark \end{cases}$$

0	0	0	0	0	0
0					
0					
0					
0					
0					

D[m, n]: optimal suffix alignment of S v.s T

Answer will be here

How to backtrace?