Slow (alignments)

- First a vote

- want to remove some topics
  - who is interested in the practical aspects of molecular dynamics simulations ?
  - who cares about simulated annealing ?

- who is interested in
  - simulations in non-physical spaces ?
  - cunning heuristics for NP-complete problems

Alignment Problem

```
. . . A N D R A N D Q R E W . . .
. . . . . . . . A N D R E W . . .
```

- or

```
. . . A N D R A N D Q R E W . . .
. . . A N D R E W . . . . . .
```

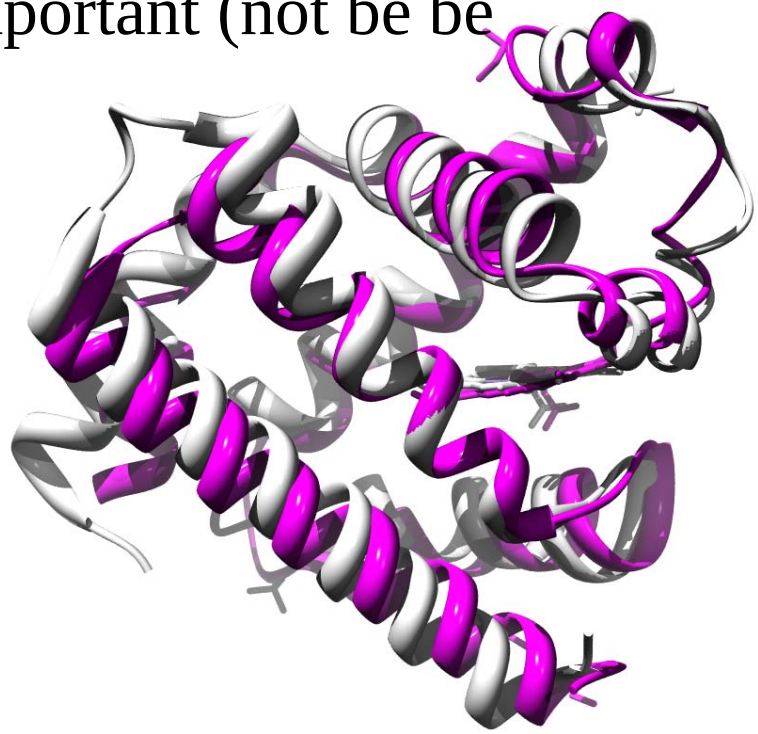
- or

```
. . . A N D R A N D Q R E W . . .
. . . . . . . . A N D - R E W . . .
```

- why ?

Who uses alignments

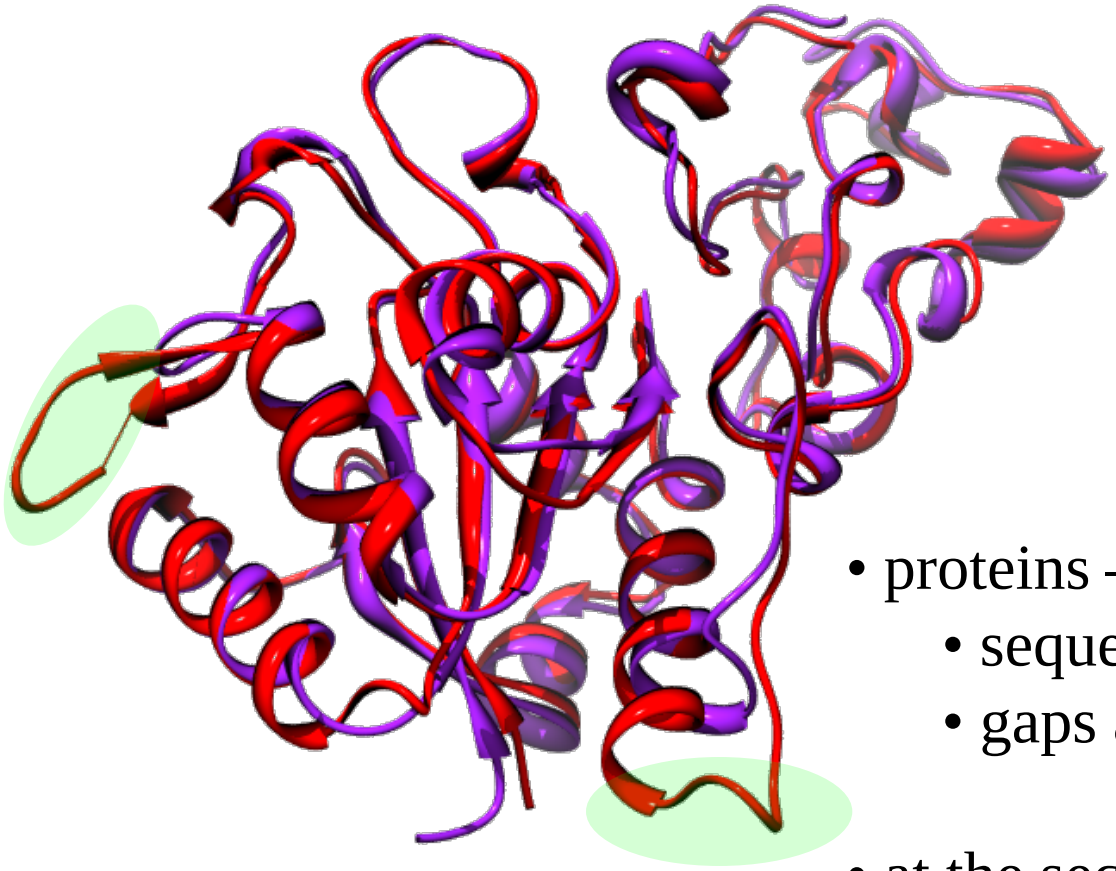
- How related are two genomes ?
- Which parts of a sequence are common to different organisms ?
- Which residues in a protein are important (not be be mutated for binding)



- tuna and horse myoglobin

- in this problem
  - we do not know the sequences

More difficult alignment



- hydrogenases
  - 40 % sequence identity
  - 2frvG & 1cc1S

- proteins – obviously similar
  - sequence identity OK
  - gaps and insertions

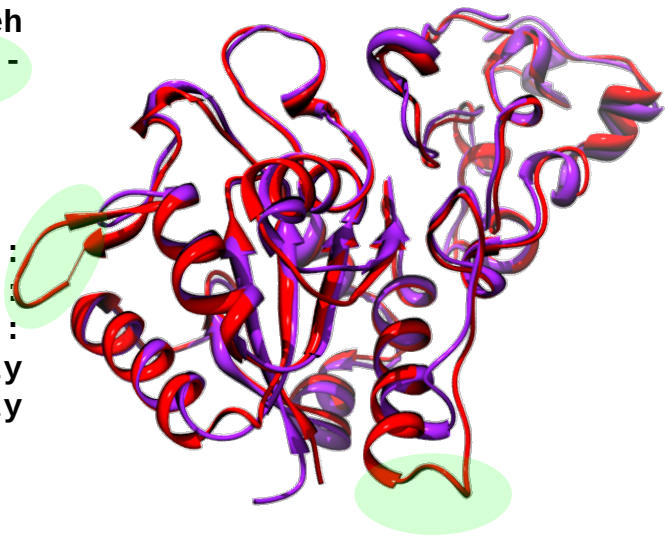
- at the sequence level ?

```
Seq ID 40.6 % (103 / 254) in 280 total including gaps
    :    1    :    2    :    3    :    4    :    5    :    6
    :    0    :    0    :    0    :    0    :    0    :    0
kkapviwvqgqgctgcsvsllnavhprikeilldvislefhptvmasegemalahmyeia
krpsvvylhnaectgcsesvlrtvdpyvdelildvismdyhetlmagaghaveea-l-he
   :    1    :    2    :    3    :    4    :    5    :
   :    0    :    0    :    0    :    0    :    0    :

    :    0    :    0    :    0    :    1    :    1    :    1
    :    7    :    8    :    9    :    0    :    1    :    2
    :    0    :    0    :    0    :    0    :    0    :    0
ekfngnfflvegaiptakegrycivgeakahhhevtmmelirdlapkslatvavgtcsa
aikg-dfvcvieggipmgdggywgk-------vggrnmydicaevapkakaviaigtcat
0     :    0    :    0          :    0    :    1    :    1
6     :    7    :    8          :    9    :    0    :    1
0     :    0    :    0          :    0    :    0    :    0

    :    1    :    1    :    1    :    1    :    1    :    1
    :    3    :    4    :    5    :    6    :    7    :    8
    :    0    :    0    :    0    :    0    :    0    :    0
yggipaaegnvtgsksvrdffadekiekllvnvpgcpphpdwmvgtlvaawshvlnpteh
yggvqaakpnptgtvgvnealgklgvkai--niagcppnpmnfvgtv--vhlltk-----
   :    1    :    1    :    1      :    1    :      1
   :    2    :    3    :    4      :    5    :      6
   :    0    :    0    :    0      :    0    :      0

    :    1    :    2    :    2    :    2         :    2    :
    :    9    :    0    :    1    :    2         :    3    :
    :    0    :    0    :    0    :    0         :    0    :
plpeldddgrpllffgdnihencpyldkydnsefaetftkpg-----ckaelgckgpsty
gmpeldkqgrpvmffgetvhdncprlkhfeagefatsfgspeakkgyclyelgckgpdty
  :    1    :    1    :    1    :    2    :    2    :    2
  :    7    :    8    :    9    :    0    :    1    :    2
  :    0    :    0    :    0    :    0    :    0    :    0
```



How big is the problem ?

- For two sequences of length 10, how many alignments ?

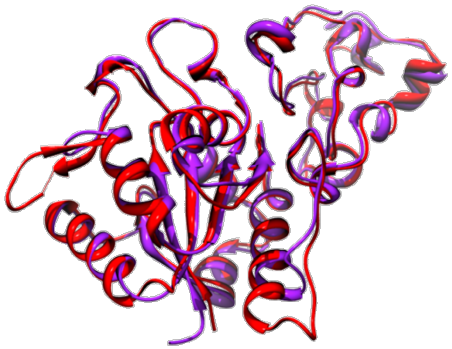
```
. . . . . . . . A B C D E F G H I J . . . .
                Q R S T U V W X Y Z
. . . . . . Q R S T U V W X Y Z + more
```

with gaps

```
              Q R S T U V W X Y – Z
              Q R S T U V W X – Y Z  then with gap 2
            Q R S T U V W X Y - - Z
                ...
```

- then with multiple gaps … combinatorial explosion
- do not tackle the problem directly

Framework



- Belief
  - there is a correct alignment
- DNA *vs* Protein – basic algorithm the same
- Evolutionary basis
  - bases (DNA) mutate occasionally
  - residues (proteins) change less often
  - gaps introduced rarely, but must be treated (protein example)
- Allow gaps/insertions, but not too many

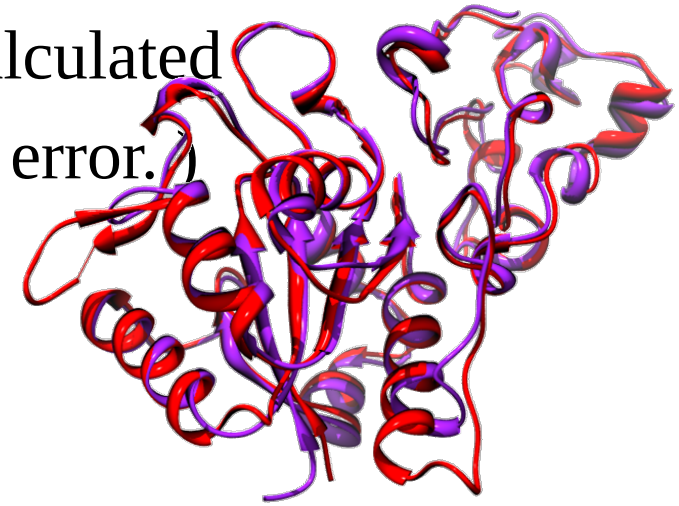
Different kinds of answers

- Define problem
  - Given some scheme for calculating alignment scores (with gaps) find alignment that maximises the score
- Remember my claims about non-optimality ?

Needleman & Wunsch / Smith Waterman	O(n^3)	first versions
Gotoh (N&W/ S&W)	O(n^2)	tiny limitation – this lecture
seeded and suffix tree methods	< O(mn)	Dr Pavlović-Lažetić

Different alignment methods

- Searching a database
  - non-redundant proteins, 1.2×10^7 sequences, 3×10^9 residues
  - cannot do 10^7 expensive alignments
  - tolerate small errors
- building a model for protein
  - only one or few alignments to be calculated
  - details of alignment important (4 Å error.)



Scoring for DNA

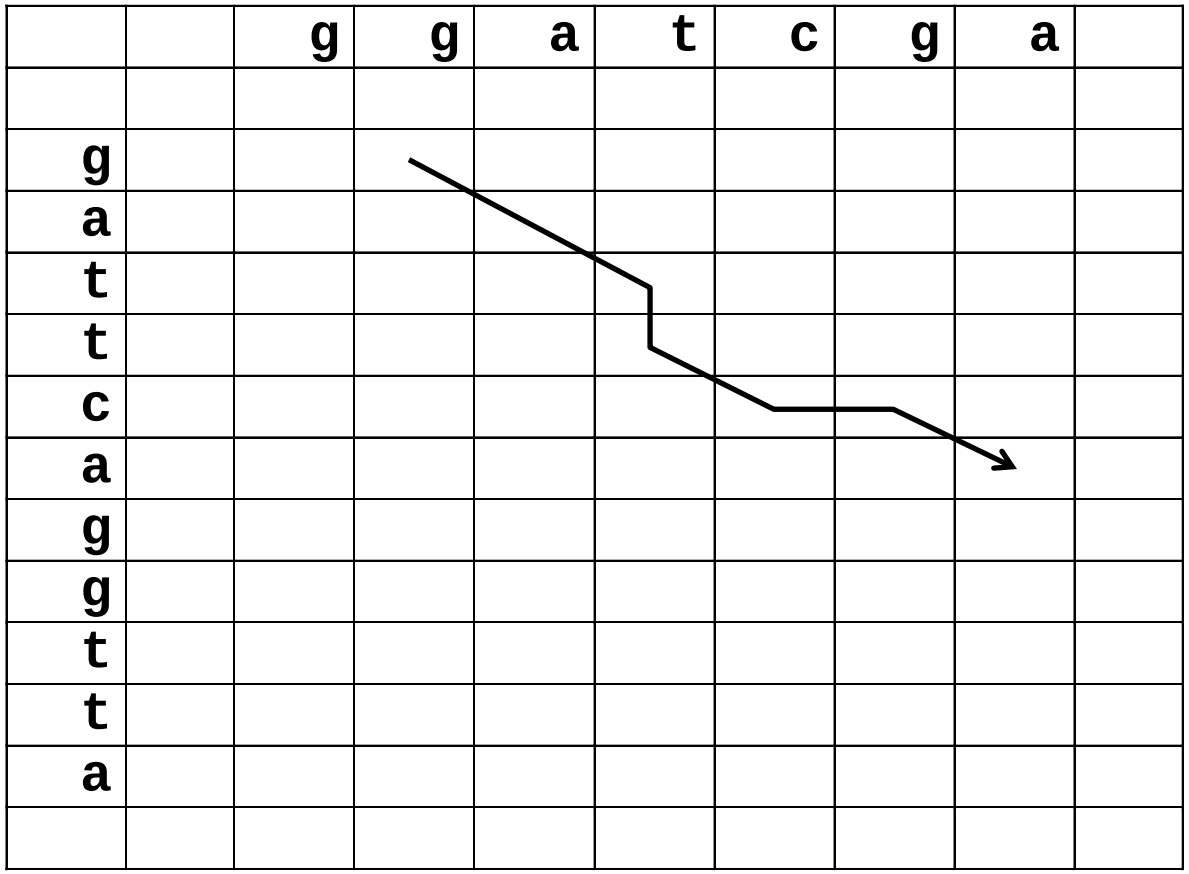
- Sensible scheme
  - matched pairs 2
  - mismatch −3
  - gaps −2

```
A C T G  - A  T T C G A
A C - G  C A  - T C T A
2 2-2 2 -2 2 -2 2 2-3 2
```

- more sophisticated..
  - gap opening costs − 2
  - gap widening costs − 1
  - so $cost = cost_{open} + (n_{gap} - 1)cost_{widen}$

Representing alignments

- sequences GATTCAGGTTA and GGATCGA

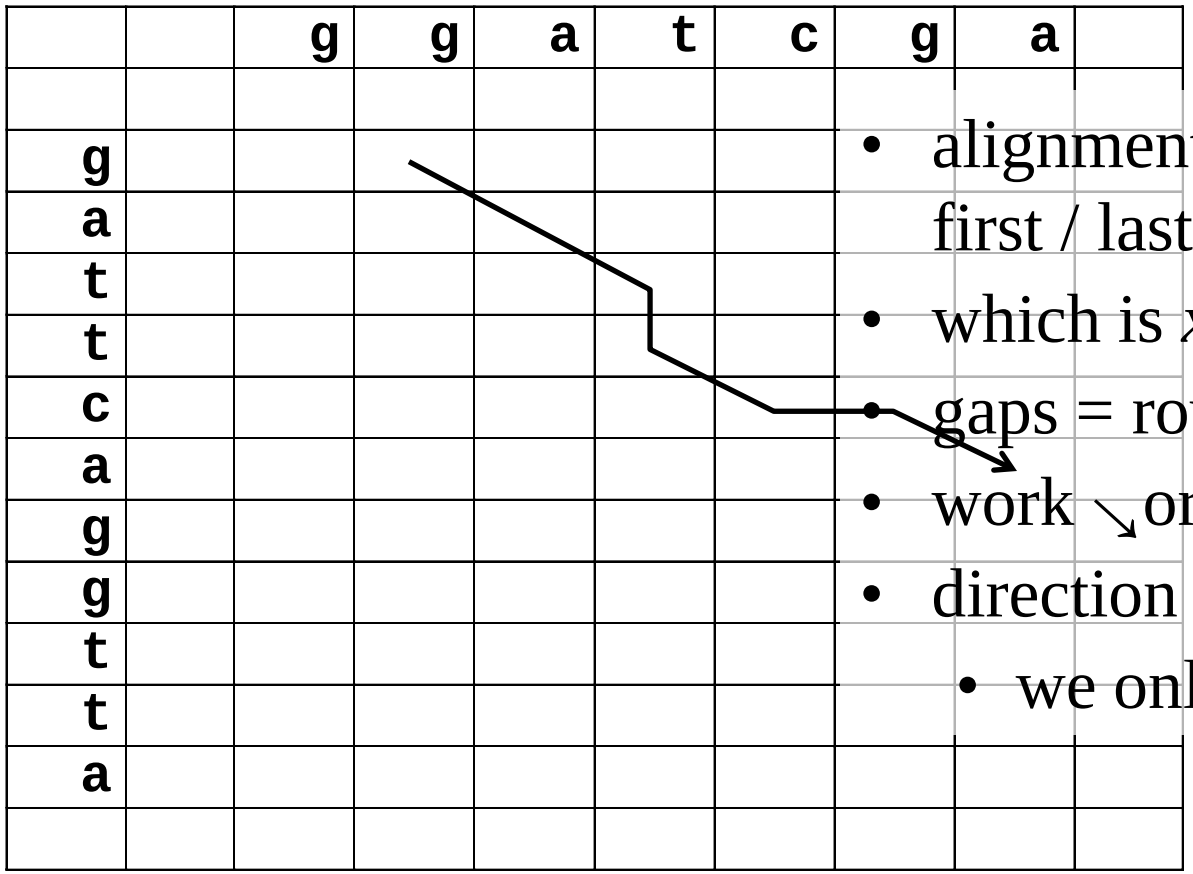


- would mean
  GGAT-CGA-----
  -GATTC-AGGTTA

- notes…

Representing alignments

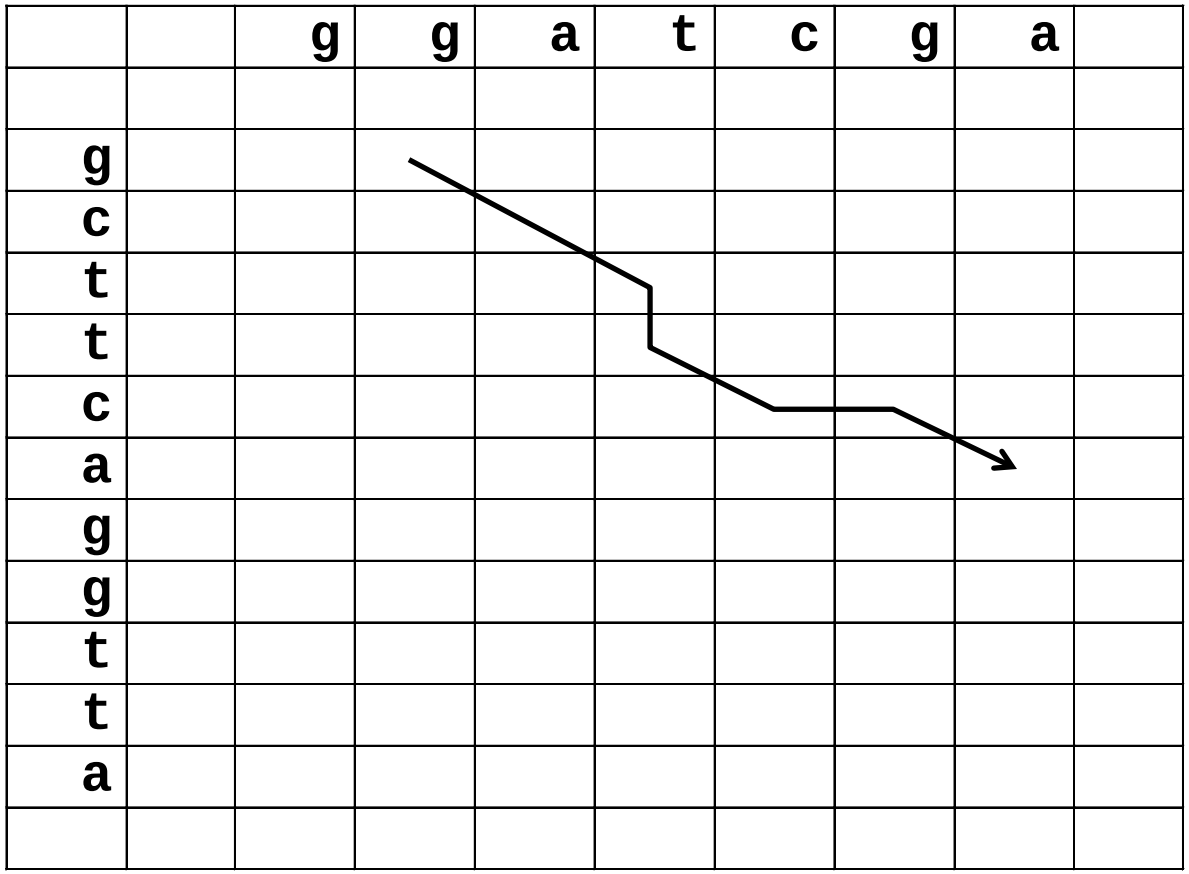
```
GGAT-CGA-----
-GATTC-AGGTTA
```



- alignment does not have to go to first / last row or column
- which is *x* and *y* is arbitrary
- gaps = row or column is skipped
- work ↘ or ↖ does not matter
- direction must be consistent
  - we only go → ↓ ↘

Representing alignments with a mismatch

- sequences GCTTCAGGTTA and GGATCGA



- would mean
  GGAT-CGA-----
  -GCTTC-AGGTTA

Calculating alignment - steps

Needleman and Wunsch algorithm

1. fill score matrix
2. find best score possible in each cell
3. traceback

fill score matrix

- For convenience, add some zeroes to the ends

		g	g	a	t	c	g	a	
	0	0	0	0	0	0	0	0	0
g	0								0
a	0								0
t	0								0
t	0								0
c	0								0
a	0								0
g	0								0
g	0								0
t	0								0
t	0								0
a	0								0
	0	0	0	0	0	0	0	0	0

Mission

- find path through this matrix with best score
- account for gaps

fill score matrix

- For convenience, add some zeroes to the ends
- Add in match, mismatch scores

		g	g	a	t	c	g	a	
	0	0	0	0	0	0	0	0	0
g	0	2	2	-3	-3	-3	2	-3	0
a	0	-3	-3	2	-3	-3	-3	2	0
t	0	-3	-3	-3	2	-3	-3	-3	0
t	0	-3	-3	-3	2	-3	-3	-3	0
c	0	-3	-3	-3	-3	2	-3	-3	0
a	0	-3	-3	2	-3	-3	2	2	0
g	0	2	2	-3	-3	-3	2	-3	0
g	0	2	2	-3	-3	-3	2	-3	0
t	0	-3	-3	-3	2	-3	-3	2	0
t	0	-3	-3	-3	2	-3	-3	-3	0
a	0	-3	-3	2	-3	-3	-3	2	0
	0	0	0	0	0	0	0	0	0

Mission

- find path through this matrix with best score
- account for gaps

Summing the elements

- start at top left
- move right, then next line
- at each cell
  - find best score it could possibly have

		g	g	a	t	c	g	a	
	0	0	0	0	0	0	0	0	0
g	0	2	2	-3	-3	-3	2	-3	0
a	0	-3	-1	4	-3	-4	-5	4	0
t	0	-3	-3	-3	6	-1	-2	-3	4
t	0	-3	-4	-4	4	3	1	0	2
c	0	-3	-5	-5	-2	6	0	-2	1
a	0	-3	-5	-6	-3	0	3	6	3
g	0	2	0	-6	-4	-1	6	0	6
g	0	2	4	-3	-4	-2	5	3	4
t	0	-3	-1	1	4	-2	-1	2	3
t	0	-3	-3	-1	3	1	-1	0	2
a	0	-3	-4	3	-4	0	-2	4	0
	0	0	-2	0	3	1	0	1	4

Diagonal (no gaps)

for each cell, 3 possible scores

1. **diagonal (no gap)**
2. best from preceding column
3. best from preceding row

		g	g	a	t	c	g	a	
	0	0	0	0	0	0	0	0	0
g	0	2	2	-3	-3	-3	2	-3	0
a	0	-3	-1	4	-3	-4	-5	4	0
t	0	-3	-3	-3	6	-1	-2	-3	4
t	0	-3	-4	-4	4	3	1	0	2
c	0	-3	-5	-5	-2	6	0	-2	1
a	0	-3	-5	-6	-3	0	3	6	3
g	0	2	0	-6	-4	-1	6	0	6
g	0	2	4	-3	-4	-2	5	3	4
t	0	-3	-1	1	4	-2	-1	2	3
t	0	-3	-3	-1	3	1	-1	0	2
a	0	-3	-4	3	-4	0	-2	4	0
	0	0	-2	0	3	1	0	1	4

GAT
GAT

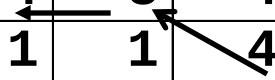
GG
GG

preceding row (gap)

for each cell, 3 possible scores

1. diagonal (no gap)
2. **best from preceding row**
3. best from preceding column

		g	g	a	t	c	g	a	
	0	0	0	0	0	0	0	0	0
g	0	2	2	-3	-3	-3	2	-3	0
a	0	-3	-1	4	-3	-4	-5	4	0
t	0	-3	-3	-3	6	-1	-2	-3	4
t	0	-3	-4	-4	4	3	1	0	2
c	0	-3	-5	-5	-2	6	0	-2	1
a	0	-3	-5	-6	-3	0	3	6	3
g	0	2	0	-6	-4	-1	6	0	6
g	0	2	4	-3	-4	-2	5	3	4
t	0	-3	-1	1	4	-2	-1	2	3
t	0	-3	-3	-1	3	1	-1	0	2
a	0	-3	-4	3	-4	0	-2	4	0
	0	0	-2	0	3	1	0	1	4



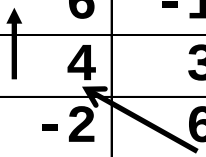
```
GAT
G-T
```

preceding column (gap)

for each cell, 3 possible scores

1. diagonal (no gap)
2. best from preceding row
3. **best from preceding column**

		g	g	a	t	c	g	a	
	0	0	0	0	0	0	0	0	0
g	0	2	2	-3	-3	-3	2	-3	0
a	0	-3	-1	4	-3	-4	-5	4	0
t	0	-3	-3	-3	6	-1	-2	-3	4
t	0	-3	-4	-4	4	3	1	0	2
c	0	-3	-5	-5	-2	6	0	-2	1
a	0	-3	-5	-6	-3	0	3	6	3
g	0	2	0	-6	-4	-1	6	0	6
g	0	2	4	-3	-4	-2	5	3	4
t	0	-3	-1	1	4	-2	-1	2	3
t	0	-3	-3	-1	3	1	-1	0	2
a	0	-3	-4	3	-4	0	-2	4	0
	0	0	-2	0	3	1	0	1	4



```
T-C
TTC
```

The order of cells

- start at top left
- every cell has best score considering all possible routes
- at end, highest score is best path

		g	g	a	t	c	g	a	
	0	0	0	0	0	0	0	0	0
g	0	2	2	-3	-3	-3	2	-3	0
a	0	-3	-1	4	-3	-4	-5	4	0
t	0	-3	-3	-3	6	-1	-2	-3	4
t	0								
c	0								
a	0								
g	0								
g	0								
t	0								
t	0								
a	0								
	0								

- would also work if we went left and up

Reading the alignment

- find highest scoring cell (last row or column)
- how did we reach this cell ?
  - how did we reach preceding cell ?
  - …

		g	g	a	t	c	g	a	
	0	0	0	0	0	0	0	0	0
g	0	2	2	-3	-3	-3	2	-3	0
a	0	-3	-1	4	-3	-4	-5	4	0
t	0	-3	-3	-3	6	-1	-2	-3	4
t	0	-3	-4	-4	4	3	1	0	2
c	0	-3	-5	-5	-2	6	0	-2	1
a	0	-3	-5	-6	-3	0	3	6	3
g	0	2	0	-6	-4	-1	6	0	6
g	0	2	4	-3	-4	-2	5	3	4
t	0	-3	-1	1	4	-2	-1	2	3
t	0	-3	-3	-1	3	1	-1	0	2
a	0	-3	-4	3	-4	0	-2	4	0
	0	0	-2	0	3	1	0	1	4

```
GGAT-CGA
-GATTC-AGGTTA
```

Trick with traceback

- for each cell
  - how did we reach it ? What was the preceding cell ?

		g	g	a	t	c	g	a	
	0	0	0	0	0	0	0	0	0
g	0	2	2	-3	-3	-3	2	-3	0
a	0	-3	-1	4	-3	-4	-5	4	0
t	0	-3	-3	-3	6	-1	-2	-3	4
t	0	-3	-4	-4	4	3	1	0	2
c	0	-3	-5	-5	-2	6	0	-2	1
a	0	-3	-5	-6	-3	0	3	6	3
g	0	2	0	-6	-4	-1	6	0	6
g	0	2	4	-3	-4	-2	5	3	4
t	0	-3	-1	1	4	-2	-1	2	3
t	0	-3	-3	-1	3	1	-1	0	2
a	0	-3	-4	3	-4	0	-2	4	0
	0	0	-2	0	3	1	0	1	4

```
GGAT-CGA
-GATTC-AGGTTA
```

Summary (Needleman and Wunsch)

- Alignments are paths through the matrix
- There is an astronomical number of possibilities (with gaps)
- This algorithm has visited all of them and found best
- allows for gap costs of form $cost = cost_{open} + (n_{gap} - 1)cost_{widen}$
- best or only method ? wait..

Cost

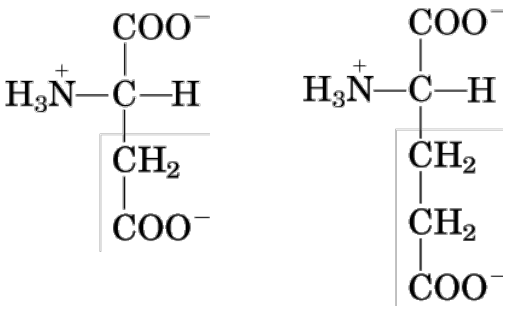
- pretend both sequences are length n
- we have to visit n^2 cells in matrix
  - each time we have to look at a row or column of length $\approx n$
- total cost n^3 or worst cost $O(n^3)$
  - in practice use $O(n^2)$ variation

Variations and Improvements

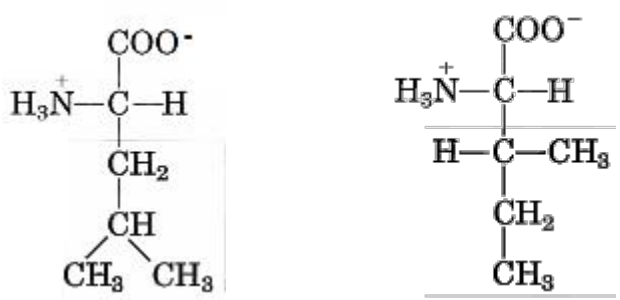
- find best scoring regions (not global optimum)
- database methods (not here)
- better scoring schemes…

Substitution Matrix

- asp and glu



- think of leu and ile



- many more similar amino acids
- glu →asp mutation, does it matter ? sometimes not
- trp →asp, big hydrophobic to small polar ? usually bad news
- relevance to alignments

Substitution Matrix

- ANDREWANDRWANDRWW aligned to QNDRDW

```
ANDREWANDRWANDRWW
QNDRDW-----------

ANDREWANDR-WANDRWW
------QNDRDW------

ANDREWANDRWANDRWW
-----------QNDRDW
```

- in our evolutionary model, only one is correct
- as chemists E and D are similar, W and D different

- how would we do it numerically ?

A serious protein similarity matrix

- blosum62:

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	4	-1	-2	-2	0	-1	-1	0	-2	-1	-1	-1	-1	-2	-1	1	0	-3	-2	0
R	-1	5	0	-2	-3	1	0	-2	0	-3	-2	2	-1	-3	-2	-1	-1	-3	-2	-3
N	-2	0	6	1	-3	0	0	0	1	-3	-3	0	-2	-3	-2	1	0	-4	-2	-3
D	-2	-2	1	6	-3	0	2	-1	-1	-3	-4	-1	-3	-3	-1	0	-1	-4	-3	-3
C	0	-3	-3	-3	9	-3	-4	-3	-3	-1	-1	-3	-1	-2	-3	-1	-1	-2	-2	-1
Q	-1	1	0	0	-3	5	2	-2	0	-3	-2	1	0	-3	-1	0	-1	-2	-1	-2
E	-1	0	0	2	-4	2	5	-2	0	-3	-3	1	-2	-3	-1	0	-1	-3	-2	-2
G	0	-2	0	-1	-3	-2	-2	6	-2	-4	-4	-2	-3	-3	-2	0	-2	-2	-3	-3
H	-2	0	1	-1	-3	0	0	-2	8	-3	-3	-1	-2	-1	-2	-1	-2	-2	2	-3
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4	2	-3	1	0	-3	-2	-1	-3	-1	3
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4	-2	2	0	-3	-2	-1	-2	-1	1
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5	-1	-3	-1	0	-1	-3	-2	-2
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5	0	-2	-1	-1	-1	-1	1
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6	-4	-2	-2	1	3	-1
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7	-1	-1	-4	-3	-2
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4	1	-3	-2	-2
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5	-2	-2	0
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11	2	-3
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	-1
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

- some features
  - diagonal
  - similar
  - different

Blosum matrix

- Many variations
  - short evolutionary times, asp → asp
  - longer times asp/glu or asn/gln very common

- using ? No change to basic algorithm

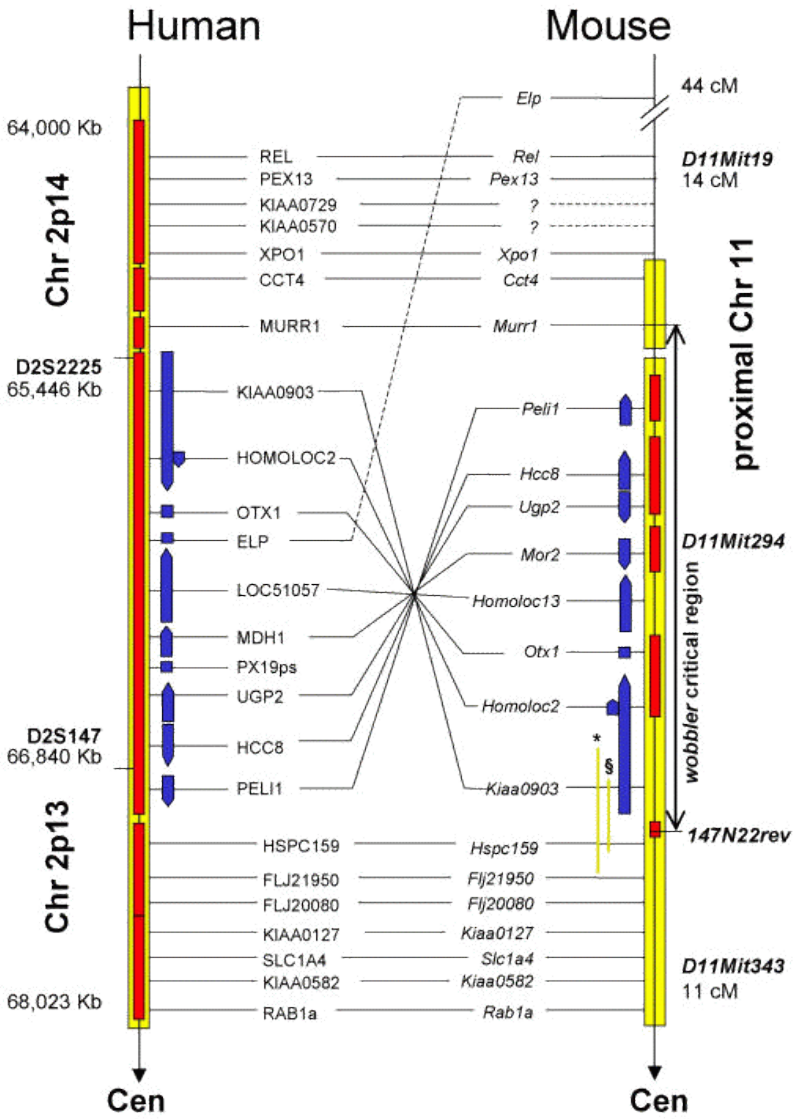
Protein vs DNA alignments

- Much DNA codes for proteins
- blocks of three bases
- CCU, CCC, CCA, CCG are all proline (3rd position degenerate)
- CCC→CCA no problem
- CCC→ACC pro → ala (you die)
  - exactly the same mutation at DNA level (C→A)

- at the DNA level, miss this effect
- at the protein level
  - synonymous mutations
  - similarity of residues

Finished with alignments ?

- Different methods
  - database searches
  - genome alignments
- Phylogeny
  - vicious approximations



synteny plot: http://home.cc.umanitoba.ca/~umlawda/39.769/presentation/presentation.html, Fristensky, B.