Another hat tossed into the million genome ring (so to speak)…

Genome wiz Venter partners with Roche in DNA sequencing deal

BY JULIE STEENHUYSEN
CHICAGO | Wed Jan 14, 2015 3:35pm EST

Tweet 23 Share 41 Share this 8+1 0 Email Print

1 OF 2 Craig Venter, founder, chairman, and president of the J. Craig Venter Institute, listens to opening statements at House Energy and Commerce hearing on 'Developments in Synthetic Genomics and Implications for Health and Energy' on Capitol Hill in Washington, in this file photo taken on May 27, 2010.
CREDIT: REUTERS/HYUNGWON KANG

RELATED NEWS
Deals of the day- Mergers and acquisitions
Roche in $1 billion-plus Foundation Medicine deal to boost cancer efforts
Rise in healthcare stocks lifts European equities

(Reuters) - Genome pioneer J. Craig Venter has signed a multi-year deal to sequence and analyze tens of thousands of genomes for Roche's Genentech unit in a deal aimed at identifying new drug targets and biomarkers, the companies said on Wednesday.

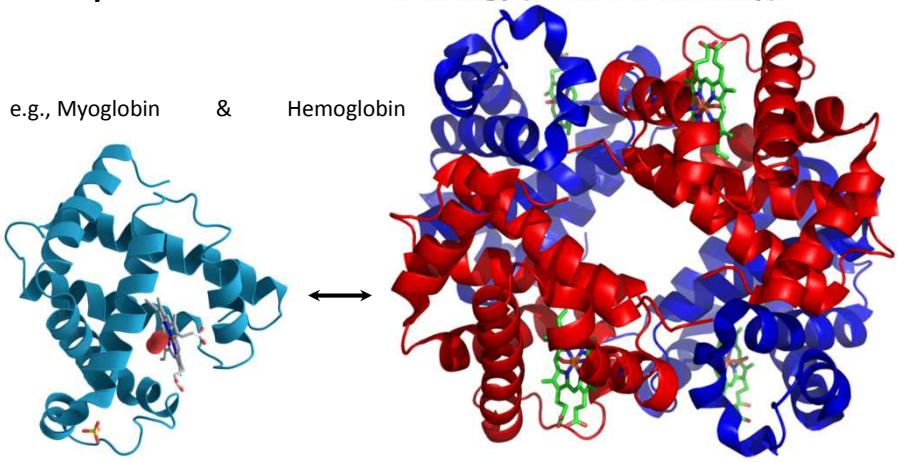
The deal is one of the biggest yet for Venter's La Jolla, California based Human Longevity Inc (HLI), a start-up formed last March with the goal of sequencing 1 million genomes by 2020. Financial terms were not disclosed.

Human Longevity Inc (HLI), a start-up formed last March with the goal of sequencing 1 million genomes by 2020.

Reuters
1-14-2015

Typically, to be "biologically related" means to <u>share a common ancestor</u>. In biology, we call this *homologous*.
Two proteins sharing a common ancestor are said to be *homologs*.
Homology often implies structural similarity & sometimes (not always) sequence similarity. **A statistically significant sequence or structural similarity can be used to infer homology (common ancestry).**

e.g., Myoglobin & Hemoglobin

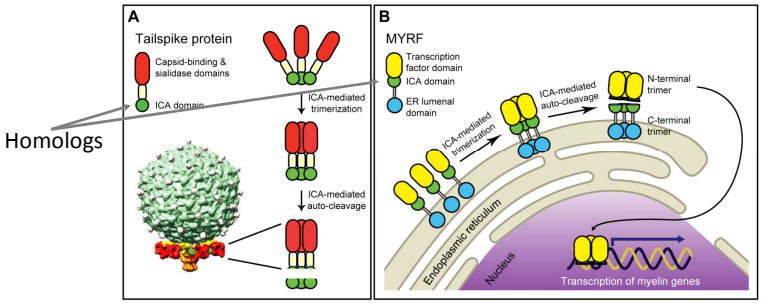


Edward Marcotte/Univ. of Texas/BCH391L/Spring 2015

http://en.wikipedia.org/wiki/File:Myoglobin.png & File:1GZX_Haemoglobin.png

In practice, searching for sequence or structural similarity is one of the most powerful computational approaches for discovering functions for genes, since we can often glean many new insights about a protein based on what is known about its homologs.

Here's an example from my own lab, where we discovered that myelinating the neurons in your brain employs the same biochemical mechanism used by bacteriophages to make capsids.
The critical breakthrough was recognizing that the human and phage proteins contained homologous domains.



Li, Park, Marcotte
PLoS Biology (2013)

Sequence alignment algorithms such as BLAST, PSI-BLAST, FASTA, and the Needleman–Wunsch & Smith-Waterman algorithms arguably comprise some of the most important driver technologies of modern biology and underlie the sequencing revolution.

So, let's start learning bioinformatics algorithms by learning how to align two protein sequences.

Live demo:

http://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastp&BLAST_PROGRAMS=blastp&PAGE_TYPE=BlastSearch&SHOW_DEFAULTS=on&LINK_LOC=blasthome

MVLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
KKVADALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFTP
AVHASLDKFLASVSTVLTSKYR

Title:All non-redundant GenBank CDS translations+PDB+SwissProt+PIR+PRF excluding environmental samples from WGS projects
Molecule Type:Protein
Update date:2015/01/18
Number of sequences:57851050

Protein sequence alignment

Two biologically related proteins with similar sequences:

```
FlgA1 EAGNVKLKRGRLDTLPPRTVLDINQLVDAISLRDLSPDQPIQLTQFRQAWRVKAGQRVNVIASGD
        ++K+K+GRLDTLPP  +L+ N    A+SLR ++  QP+     R+ W +KAGQ V V+A G+
FlgA2 TLQDIKMKQGRLDTLPPGALLEPNFAQGAVSLRQINAGQPLTRNMLRRLWIIKAGQDVQVLALGE
```

Also biologically related (& fold up into the same 3D protein structure):

```
FlgA1 EAGNVKLKRGRLDTLPPRTVLDINQLVDAISLRDLSPDQPIQLTQFRQAWRVKAGQRVNVIASGD
         A   +         P        +L   I+ R L P + I     R+AW V+ G  V V
FlgA3 LAALKQVTLIAGKHKPDAMATHAEELQGKIAKRTLLPGRYIPTAAIREAWLVEQGAAVQVFFIAG
```

But these are biologically unrelated (& fold up into unrelated structures):

```
FlgA1 AGNVKLKRGRLDTLPPRTVLDINQLVDAISLRDLSPDQPIQLTQFRQA-WRVKAGQRVNVIASGD
      AG+V  K G +  + PRT  ++      I+     P  PI    +++A WRV A + V V+  GD
HvcPP AGHV--KNGTMRIVGPRTCSNVWNGTFPINATTTGPSIPIPAPNYKKALWRVSATEYVEVVRVGD
```

(FYI, we'll draw examples from Durbin *et al.*, *Biological Sequence Analysis*, Ch. 1 & 2).

To align two sequences, we need to perform 3 steps:

1. **We need some way to decide which alignments are better than others.**
   For this, we'll invent a way to give the alignments a "score" indicating their quality.

2. **Align the two proteins so that they get the best possible score.**

3. **Decide if the score is "good enough" for us to believe the alignment is biologically significant.**

To align two sequences, we need to perform 3 steps:

1. **We need some way to decide which alignments are better than others.**
   For this, we'll invent a way to give the alignments a "score" indicating their quality.

2. **Align the two proteins so that they get the best possible score.**

3. **Decide if the score is "good enough" for us to believe the alignment is biologically significant.**

Edward Marcotte/Univ. of Texas/BCH391L/Spring 2015

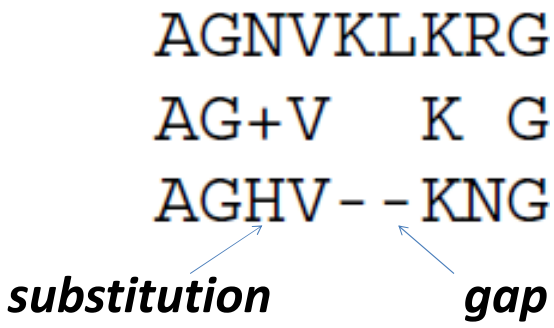
We'll treat mutations as independent events.

This allows us to create an ***additive scoring scheme.***
The score for a sequence alignment will be the sum of the scores for aligning each of the individual positions in two sequences.

What kind of mutations should we expect?

Substitutions, ***insertions*** and ***deletions***.

Insertions and deletions can be treated as equivalent events by considering one or the other sequence as the reference, and are usually called ***gaps***.



Let's consider two models:

First, a **random** model, where amino acids in the sequences occur independently at some given frequencies.

The probability of observing an alignment between *x* and *y* is just the product of the frequencies (q) with which we find each amino acid.

We can write this as:

What does the capital pi mean?

$$P(x, y \mid R) = \prod_i q_{x_i} \prod_j q_{y_j}$$

What's this mean?

What's this mean?

Second, a **match** model, where amino acids at a given position in the alignment arise from some common ancestor with a probability given by the joint probability p_{ab}.

So, under this model, the probability of the alignment is the product of the probabilities of seeing the individual amino acids aligned.

We can write that as:

What does the capital pi mean again?

$$P(x, y \mid M) = \prod_i p_{x_i y_i}$$

What's this mean?

What's this mean?

To decide which model better describes an alignment, we'll take the ratio:

$$\frac{P(x,y \mid M)}{P(x,y \mid R)} = \frac{\prod_i p_{x_i y_i}}{\prod_i q_{x_i} \prod_j q_{y_j}} = \prod_i \frac{p_{x_i y_i}}{q_{x_i} q_{y_i}}$$

What did these mean again?

Such a ratio of probabilities under 2 different models is called an ***odds ratio***.

Where else have you heard odds ratios used?

Basically: if the ratio > 1, model *M* is more probable
if < 1, model *R* is more probable.

Now, to convert this to an additive score *S*, we can simply take the logarithm of the odds ratio (called the ***log odds ratio***):

$$S = \sum_i s(x_i, y_i)$$

This is just the score for aligning one amino acid with another amino acid:

$$s(a,b) = \log\left(\frac{p_{ab}}{p_a p_b}\right)$$

Here written *a* and *b* rather than x_i and y_i to emphasize that this score reflects the *inherent preference* of the two amino acids (*a* and *b*) to be aligned.

Almost done with step 1…

The last trick:
Take a big set of pre-aligned protein sequence alignments (that are correct!) and measure all of the pairwise amino acid substitution scores (the $s(a,b)$'s). Put them in a 20x20 ***amino acid substitution matrix :***

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	5	-2	-1	-2	-1	-1	-1	0	-2	-1	-2	-1	-1	-3	-1	1	0	-3	-2	0
R	-2	7	-1	-2	-4	1	0	-3	0	-4	-3	3	-2	-3	-3	-1	-1	-3	-1	-3
N	-1	-1	7	2	-2	0	0	0	1	-3	-4	0	-2	-4	-2	1	0	-4	-2	-3
D	-2	-2	2	8	-4	0	2	-1	-1	-4	-4	-1	-4	-5	-1	0	-1	-5	-3	-4
C	-1	-4	-2	-4	13	-3	-3	-3	-3	-2	-2	-3	-2	-2	-4	-1	-1	-5	-3	-1
Q	-1	1	0	0	-3	7	2	-2	1	-3	-2	2	0	-4	-1	0	-1	-1	-1	-3
E	-1	0	0	2	-3	2	6	-3	0	-4	-3	1	-2	-3	-1	-1	-1	-3	-2	-3
G	0	-3	0	-1	-3	-2	-3	8	-2	-4	-4	-2	-3	-4	-2	0	-2	-3	-3	-4
H	-2	0	1	-1	-3	1	0	-2	10	-4	-3	0	-1	-1	-2	-1	-2	-3	2	-4
I	-1	-4	-3	-4	-2	-3	-4	-4	-4	5	2	-3	2	0	-3	-3	-1	-3	-1	4
L	-2	-3	-4	-4	-2	-2	-3	-4	-3	2	5	-3	3	1	-4	-3	-1	-2	-1	1
K	-1	3	0	-1	-3	2	1	-2	0	-3	-3	6	-2	-4	-1	0	-1	-3	-2	-3
M	-1	-2	-2	-4	-2	0	-2	-3	-1	2	3	-2	7	0	-3	-2	-1	-1	0	1
F	-3	-3	-4	-5	-2	-4	-3	-4	-1	0	1	-4	0	8	-4	-3	-2	1	4	-1
P	-1	-3	-2	-1	-4	-1	-1	-2	-2	-3	-4	-1	-3	-4	10	-1	-1	-4	-3	-3
S	1	-1	1	0	-1	0	-1	0	-1	-3	-3	0	-2	-3	-1	5	2	-4	-2	-2
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	2	5	-3	-2	0
W	-3	-3	-4	-5	-5	-1	-3	-3	-3	-3	-2	-3	-1	1	-4	-4	-3	15	2	-3
Y	-2	-1	-2	-3	-3	-1	-2	-3	2	-1	-1	-2	0	4	-3	-2	-2	2	8	-1
V	0	-3	-3	-4	-1	-3	-3	-4	-4	4	1	-3	1	-1	-3	-2	0	-3	-1	5

This is the **BLOSUM50** matrix.
(The numbers are scaled & rounded off to the nearest integer):

What's the score for aspartate (D) aligning with itself?
How about aspartate with phenylalanine (F)? Why?

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	5	-2	-1	-2	-1	-1	-1	0	-2	-1	-2	-1	-1	-3	-1	1	0	-3	-2	0
R	-2	7	-1	-2	-4	1	0	-3	0	-4	-3	3	-2	-3	-3	-1	-1	-3	-1	-3
N	-1	-1	7	2	-2	0	0	0	1	-3	-4	0	-2	-4	-2	1	0	-4	-2	-3
D	-2	-2	2	8	-4	0	2	-1	-1	-4	-4	-1	-4	-5	-1	0	-1	-5	-3	-4
C	-1	-4	-2	-4	13	-3	-3	-3	-3	-2	-2	-3	-2	-2	-4	-1	-1	-5	-3	-1
Q	-1	1	0	0	-3	7	2	-2	1	-3	-2	2	0	-4	-1	0	-1	-1	-1	-3
E	-1	0	0	2	-3	2	6	-3	0	-4	-3	1	-2	-3	-1	-1	-1	-3	-2	-3
G	0	-3	0	-1	-3	-2	-3	8	-2	-4	-4	-2	-3	-4	-2	0	-2	-3	-3	-4
H	-2	0	1	-1	-3	1	0	-2	10	-4	-3	0	-1	-1	-2	-1	-2	-3	2	-4
I	-1	-4	-3	-4	-2	-3	-4	-4	-4	5	2	-3	2	0	-3	-3	-1	-3	-1	4
L	-2	-3	-4	-4	-2	-2	-3	-4	-3	2	5	-3	3	1	-4	-3	-1	-2	-1	1
K	-1	3	0	-1	-3	2	1	-2	0	-3	-3	6	-2	-4	-1	0	-1	-3	-2	-3
M	-1	-2	-2	-4	-2	0	-2	-3	-1	2	3	-2	7	0	-3	-2	-1	-1	0	1
F	-3	-3	-4	-5	-2	-4	-3	-4	-1	0	1	-4	0	8	-4	-3	-2	1	4	-1
P	-1	-3	-2	-1	-4	-1	-1	-2	-2	-3	-4	-1	-3	-4	10	-1	-1	-4	-3	-3
S	1	-1	1	0	-1	0	-1	0	-1	-3	-3	0	-2	-3	-1	5	2	-4	-2	-2
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	2	5	-3	-2	0
W	-3	-3	-4	-5	-5	-1	-3	-3	-3	-3	-2	-3	-1	1	-4	-4	-3	15	2	-3
Y	-2	-1	-2	-3	-3	-1	-2	-3	2	-1	-1	-2	0	4	-3	-2	-2	2	8	-1
V	0	-3	-3	-4	-1	-3	-3	-4	-4	4	1	-3	1	-1	-3	-2	0	-3	-1	5

Using this matrix, **we can score any alignment as the sum of scores of individual pairs of amino acids**.

For example, the top alignment in our earlier example:

```
FlgA1 EAGNVKLKRGRLDTLPPRTVLDINQLVDAISLRDLSPDQPIQLTQFRQAWRVKAGQRVNVIASGD
         ++K+K+GRLDTLPP  +L+ N    A+SLR ++  QP+     R+ W +KAGQ V V+A G+
FlgA2 TLQDIKMKQGRLDTLPPGALLEPNFAQGAVSLRQINAGQPLTRNMLRRLWIIKAGQDVQVLALGE
```

gets the score:

S(FlgA1,FlgA2) = – 1 – 2 – 2 + 2 + 4 + 6 + ... = 186

We also need to penalize **gaps**. For now, let's just use a constant penalty ***d*** for each amino acid gap in an alignment, *i. e.*:

the penalty for a gap of length g = -g*d

PAM vs. BLOSUM

Margaret Dayhoff (1925-1983)
Developed point accepted mutation matrices (PAM matrices)

Steve and Jorja Henikoff
Developed BLOSUM matrices

Calibrated for different evolutionary times
PAM-*n* = *n* substitutions per 100 residues
e.g. matrices from PAM1 to PAM250 measure PAM1, calculate higher PAMs from that

Explicit model of evolution
(calculated using a phylogenetic tree)

Calibrated for different % identity sequences
BLOSUM-*n* = for sequences of about *n* % identity
averages substitution probabilities over sequence clusters, gives better estimates for highly divergent cases

Implicit model of evolution
(calculated from blocks of aligned sequences)

To align two sequences, we need to perform 3 steps:

1. We need some way to decide which alignments are better than others.
   For this, we'll invent a way to give the alignments a "score" indicating their quality.
2. **Align the two proteins so that they get the best possible score.**
3. Decide if the score is "good enough" for us to believe the alignment is biologically significant.

We'll use something called ***dynamic programming.***

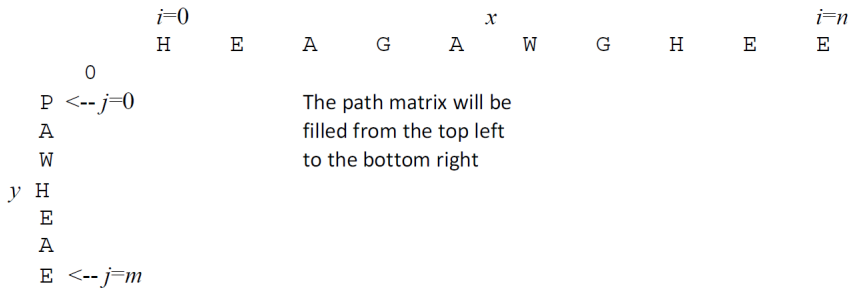
This is **mathematically guaranteed** to find the best scoring alignment, and uses ***recursion.*** This means problems are broken into sub-problems, which are in turn broken into sub-problems, etc, until the simplest sub-problems can be solved.

We're going to find the best ***local*** alignment—the best matching internal alignment—without forcing all of the amino acids to align (i.e. to match ***globally***).

i.e., this → ATGCAT
ATGCAT

Not this → ACGTTATGCATGACGTA
-C---ATGCAT----T-

Here's the main idea:
We'll make a ***path matrix***, showing the possible alignments and their scores. There are simple rules for how to fill in the matrix. ***This will test all possible alignments & give us the top-scoring alignment between the two sequences***.



Here are the rules:
For a given square in the matrix $F(i,j)$, we look at the squares to its left $F(i\text{-}1,j)$, top $F(i,j\text{-}1)$, and top-left $F(i\text{-}1,j\text{-}1)$. Each should have a score.

We consider **3 possible events** & **choose the one scoring the highest**:

(1) x_i is aligned to y_j $\quad F(i\text{-}1,j\text{-}1) + s(x_i,y_j)$

(2) x_i is aligned to a gap $\quad F(i\text{-}1,j) - d$

(3) y_j is aligned to a gap $\quad F(i,j\text{-}1) - d$

For this example, we'll use $d = 8$. We also set the left-most & top-most entries to zero.

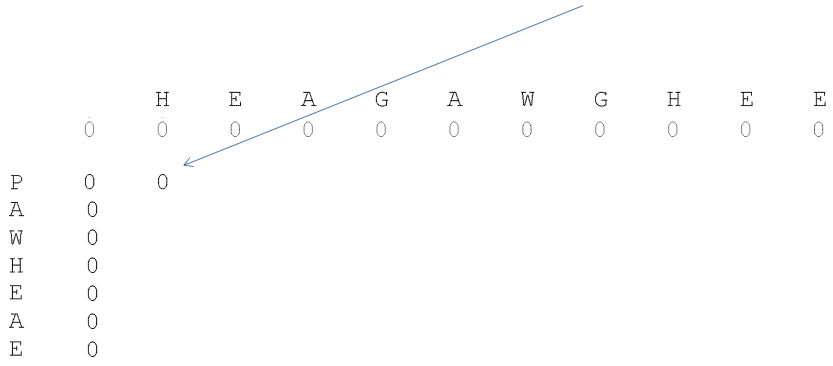
Just two more rules:

If the score is negative, set it equal to zero.

At each step, we also keep track of which event was chosen by **drawing an arrow from the cell we just filled back to the cell which contributed its score to this one.**

That's it! Just repeat this to fill the entire matrix.

Here we go! Start with the borders & the first entry.



		H	E	A	G	A	W	G	H	E	E
	0	0	0	0	0	0	0	0	0	0	0
P	0	0									
A	0										
W	0										
H	0										
E	0										
A	0										
E	0										

Why is this zero?
What's the score from our BLOSSUM matrix for substituting H for P?

Next round!

		H	E	A	G	A	W	G	H	E	E
	0	0	0	0	0	0	0	0	0	0	0
P	0	0	0								
A	0	0	0								
W	0										
H	0										
E	0										
A	0										
E	0										

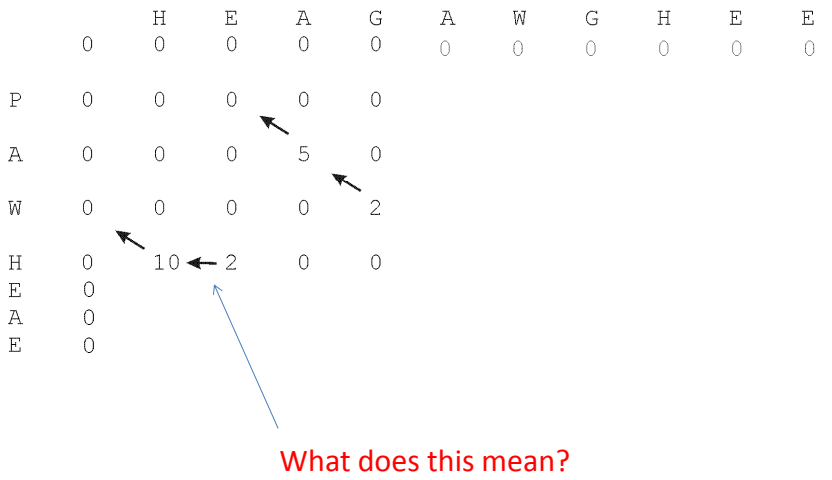
Terrible! Again, none of the possible give positive scores.
We have to go a bit further in before we find a positive score…

A few more rounds, and a positive score at last!

		H	E	A	G	A	W	G	H	E	E
	0	0	0	0	0	0	0	0	0	0	0
P	0	0	0	0							
A	0	0	0	5							
W	0	0	0								
H	0										
E	0										
A	0										
E	0										

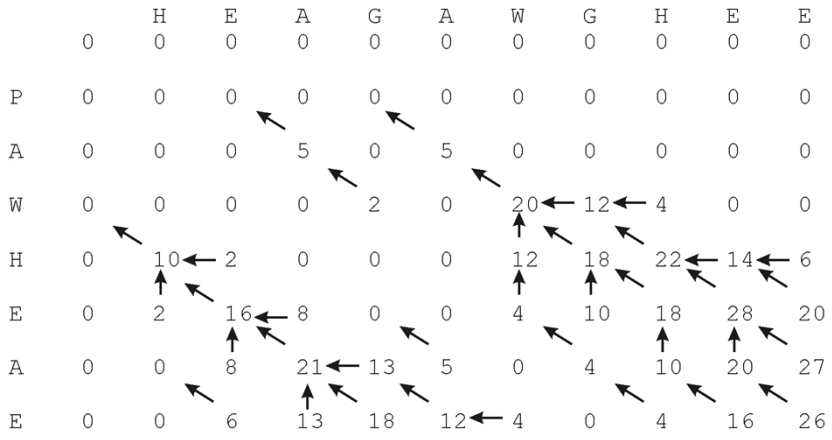
How did we get this one?

& a few more rounds…



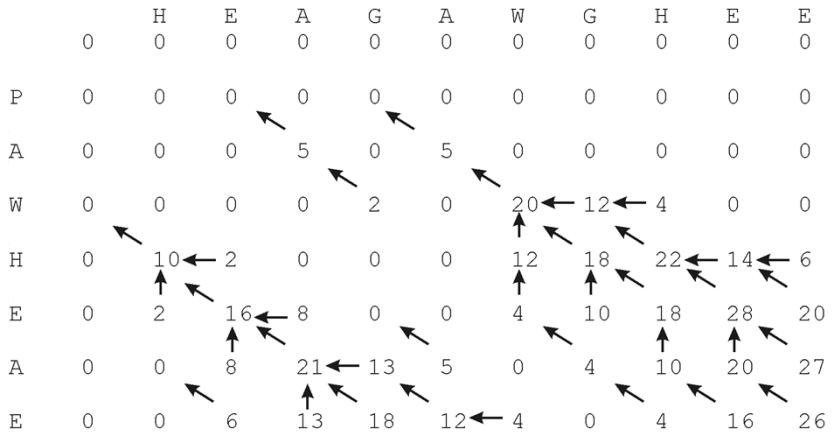
What does this mean?

The whole thing filled in!



Now, find the optimal alignment using a ***traceback*** process:

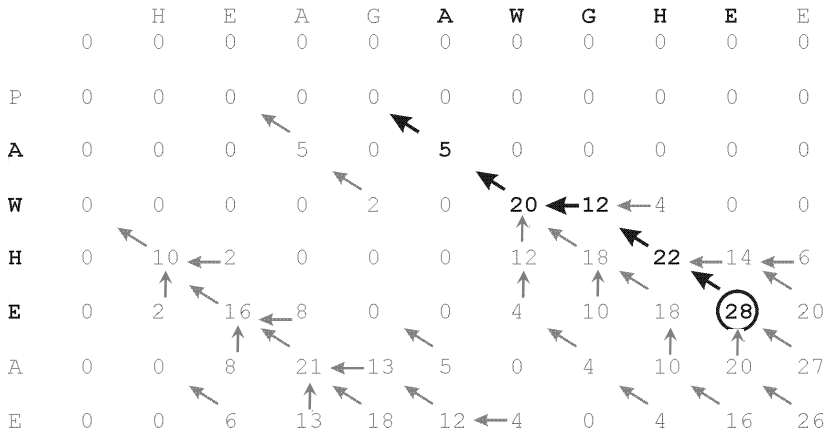
Look for the highest score, then follow the arrows back.
The alignment "grows" from right to left



		H	E	A	G	A	W	G	H	E	E
	0	0	0	0	0	0	0	0	0	0	0
P	0	0	0	0	0	0	0	0	0	0	0
A	0	0	0	5	0	5	0	0	0	0	0
W	0	0	0	0	2	0	20	12	4	0	0
H	0	10	2	0	0	0	12	18	22	14	6
E	0	2	16	8	0	0	4	10	18	28	20
A	0	0	8	21	13	5	0	4	10	20	27
E	0	0	6	13	18	12	4	0	4	16	26

This gives the following alignment: AWGHE
AW-HE

(Note: for gaps, the arrow points to the sequence that gets the gap)



		H	E	A	G	A	W	G	H	E	E
	0	0	0	0	0	0	0	0	0	0	0
P	0	0	0	0	0	0	0	0	0	0	0
A	0	0	0	5	0	5	0	0	0	0	0
W	0	0	0	0	2	0	20	12	4	0	0
H	0	10	2	0	0	0	12	18	22	14	6
E	0	2	16	8	0	0	4	10	18	28	20
A	0	0	8	21	13	5	0	4	10	20	27
E	0	0	6	13	18	12	4	0	4	16	26

To align two sequences, we need to perform 3 steps:

1. We need some way to decide which alignments are better than others.
   For this, we'll invent a way to give the alignments a "score" indicating their quality.

2. Align the two proteins so that they get the best possible score.

3. **Decide if the score is "good enough" for us to believe the alignment is biologically significant.**

This algorithm always gives the best alignment.

Every pair of sequences can be aligned in some fashion.

So, when is a score "good enough"?

How can we figure this out?

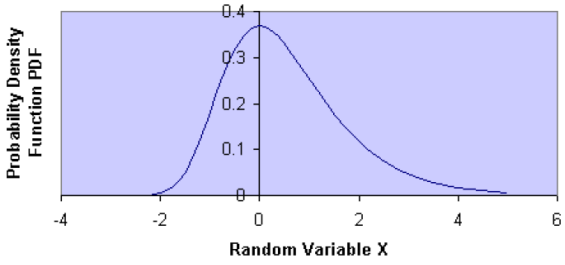
Here's one approach:

Shuffle one sequence. Calculate the best alignment & its score. Repeat 1000 times.

If we never see a score as high as the real one, we say the real score has <1 in a 1000 chance of happening just by luck.

But if we want something that only occurs < 1 in a million, we'd have to shuffle 1,000,000 times…

Luckily, alignment scores follow a well-behaved distribution, the ***extreme value distribution***, so we can do a few trials & fit to this.



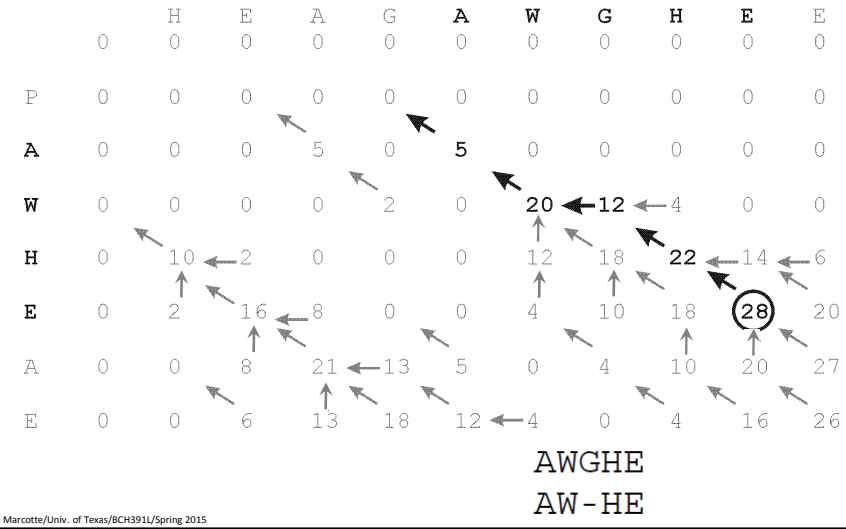
random trials & their average score

$$p(\text{max score} \leq X) \approx e^{-kNe^{\lambda(X-\mu)}}$$

This p-value gives the significance of your alignment. But, if we search a database and perform many alignments, we still need something more (next time).

Describe the shape & can be fit from a few trials

Some extensions: Local vs. global alignments
How might you force the full sequences to align?



Some extensions: Local vs. global alignments
How might you force the full sequences to align?

A few tiny changes:

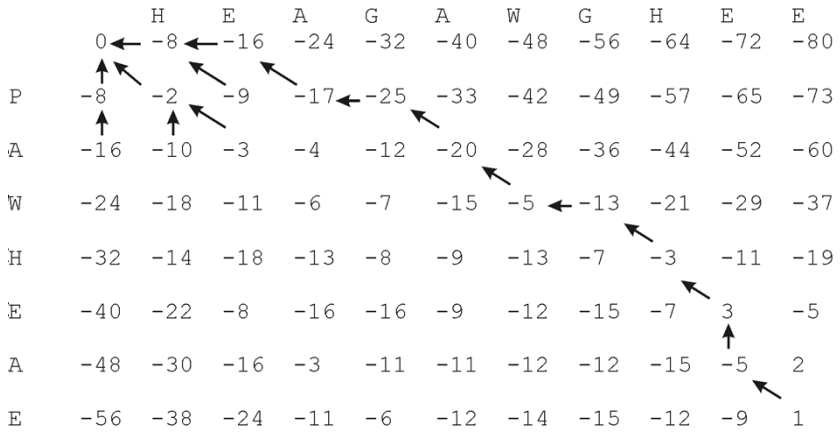
Initialize only the top left cell of the path matrix to zero
(not all top and left cells).

Leave the negative values (don't set them to zero).

The optimal alignment should start at the top left cell and finish at the bottom right cell of the path matrix.

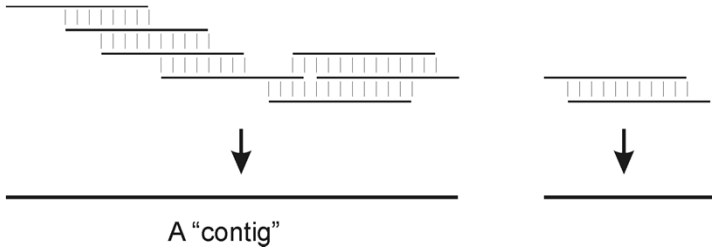
Start the trace-back at the bottom right cell

Some extensions: Local vs. global alignments
How might you force the full sequences to align?



HEAGAWGHE-E
--P-AW-HEAE

Some extensions:
What about overlapping sequences?
e.g. as in 'shotgun sequencing' genomes where 'contigs' are built up from overlapping sequences



Some extensions:
What about overlapping sequences?

Modify global alignment to not penalize overhangs:

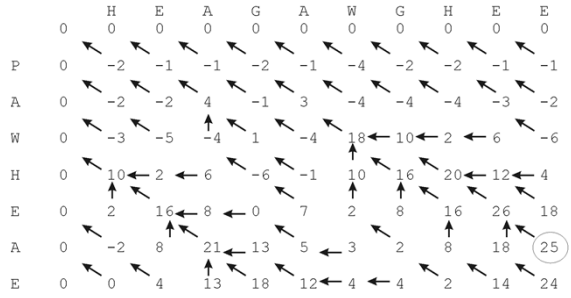
The optimal alignment should start at the top or left edge and finish at the bottom or right edge of the path matrix.

Set these boundary conditions :

$F(i,0) = 0$ for i=1 to n
$F(0,j) = 0$ for j=1 to m

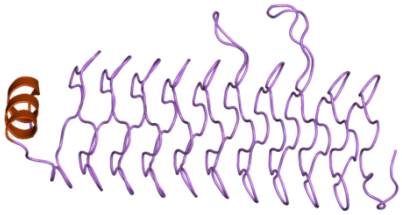
Start the traceback at the cell with the highest score on the right or bottom border

Some extensions:
What about overlapping sequences?
e.g. as in 'shotgun sequencing' genomes where 'contigs' are built up from overlapping sequences



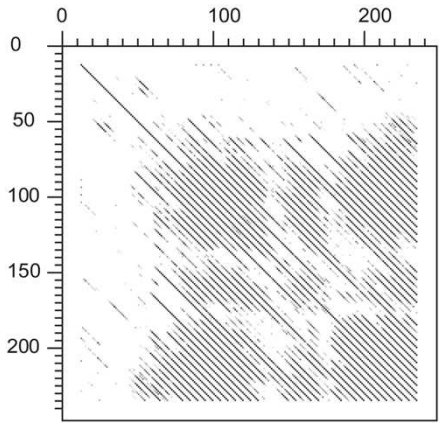
(overhang = HEA) **GAWGHEE**
PAW-HEA **(overhang = E)**

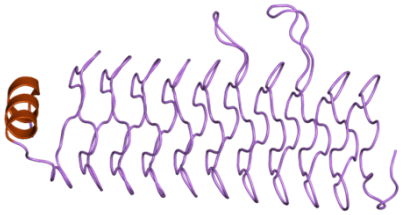
Some extensions:
How might you find repetitive sequences?



Structure of the pentapeptide repeat protein HetL
(from wiki, PMID18952182)

Align the sequence to itself and ignore the diagonal (optimal) alignment
→ High-scoring off-diagonal alignments will be repeats





Structure of the pentapeptide repeat protein HetL
(from wiki, PMID18952182)

Dot plot (quick visualization of sequence similarity) of the pentapeptide repeat protein HglK protein vs. itself
(http://en.wikipedia.org/wiki/Pentapeptide_repeat)