What is DNA?

(An overly simple, yet sufficient, answer)

1. DNA contains the instructions needed to construct other components of cells such as *protein* and *RNA*.

What is DNA?

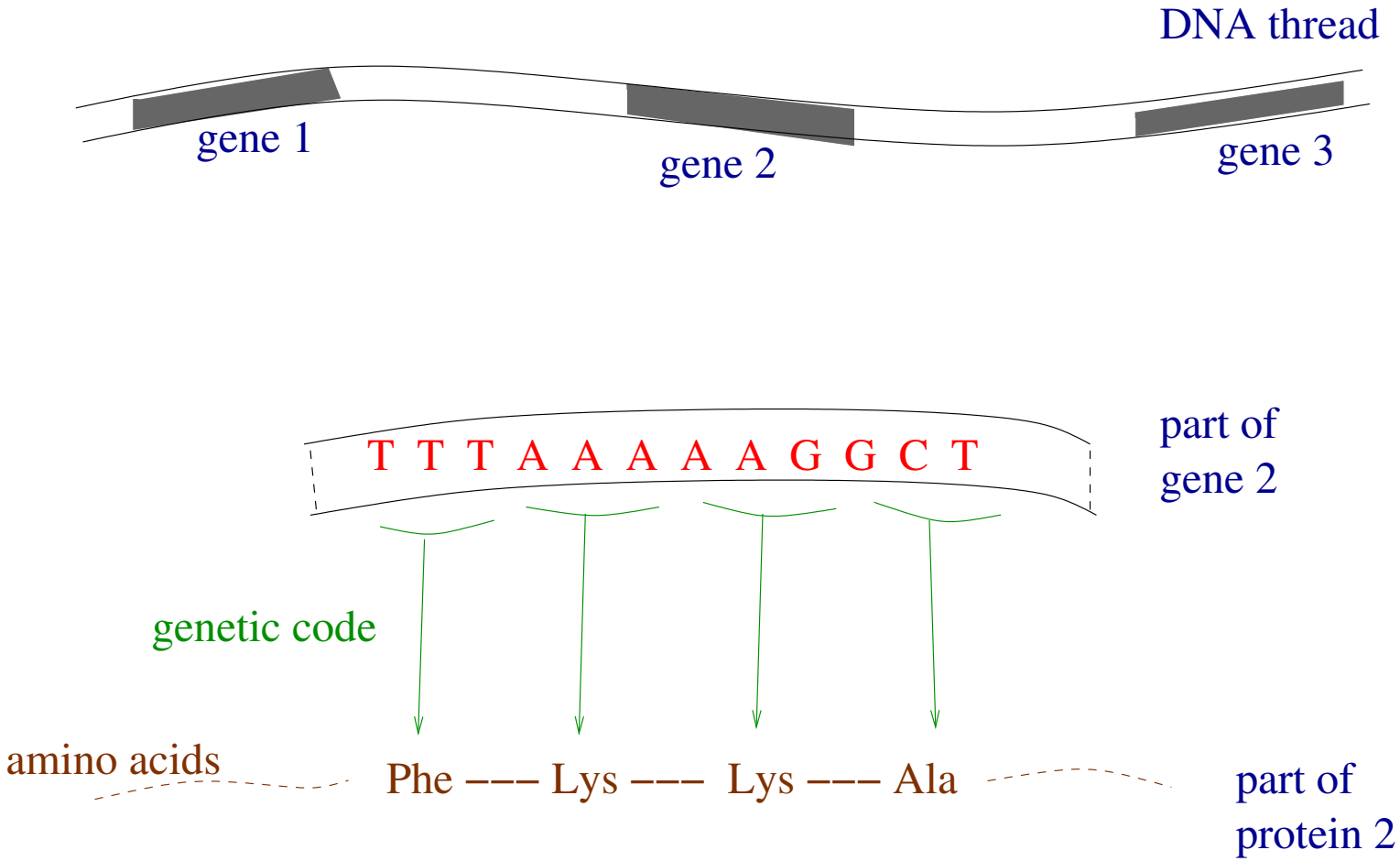
(An overly simple, yet sufficient, answer)

1. DNA contains the instructions needed to construct other components of cells such as *protein* and *RNA*.
2. There are 20 different kinds of *amino acid* that combine to make proteins. The cell DNA tells a cell the order in which to assemble the amino acids, and the length to be assembled.

What is DNA?

(An overly simple, yet sufficient, answer)

DNA is a string of four different nucleotides: **A**denine, **G**uanine, **C**ytosine and **T**hymine.



In 1984 the newly sequenced cancer causing ν-sys gene was compared against all known genes. It was 'similar' to a normal gene involved in growth. Suddently it became clear that cancer might be caused by a growth gene being switched on at the wrong time.

Question: How similar are

$$V = \quad A\ T\ C\ T\ G\ A\ T$$

and

$$W = \quad T\ G\ C\ A\ T\ A \qquad ?$$

Question: How similar are

$$V = \quad A\ T\ C\ T\ G\ A\ T$$

and

$$W = \quad T\ G\ C\ A\ T\ A \qquad ?$$

One measure of similarity is the length of a longest common subsequence (LCS) which we denote by $s(V, W)$.

Question: How similar are

$$V = \quad A\ T\ C\ T\ G\ A\ T$$

and

$$W = \quad T\ G\ C\ A\ T\ A \qquad ?$$

One measure of similarity is the length of a longest common subsequence (LCS) which we denote by $s(V, W)$.

We have $v_2 v_3 v_4 v_6 = TCTA$ and $w_1 w_3 w_5 w_6 = TCTA$ and there is no longer LCS, so

$$s(V, W) = 4.$$

A more general measure is based on sequence alignment and a choice of constants μ, ρ.

A more general measure is based on sequence alignment and a choice of constants μ, ρ.

For

$$V = ATCTGAT \quad \text{and} \quad W = TGCATA$$

we have an *alignment*

$$\begin{array}{cccccccccc} v_1 v_2 - v_3 - v_4 v_5 v_6 v_7 & = & A & T & - & C & - & T & G & A & T \\ -w_1 w_2 w_3 w_4 w_5 - w_6 - & = & - & T & G & C & A & T & - & A & - \end{array}$$

Column Score

Define the *score* of a column in an alignment to be

$$\delta\begin{pmatrix} x \\ y \end{pmatrix} = 1 \quad \text{if } x = y, \quad \text{(Match)}$$

$$\delta\begin{pmatrix} x \\ y \end{pmatrix} = -\mu \quad \text{if } x \neq y, \quad \text{(Mismatch)}$$

$$\delta\begin{pmatrix} x \\ - \end{pmatrix} = \delta\begin{pmatrix} - \\ y \end{pmatrix} = -\rho \quad \text{if } x \neq y \quad \text{(Deletion/Insertion)}.$$

Column Score

Define the *score* of a column in an alignment to be

$$\delta\begin{pmatrix} x \\ y \end{pmatrix} = 1 \quad \text{if } x = y, \quad \text{(Match)}$$

$$\delta\begin{pmatrix} x \\ y \end{pmatrix} = -\mu \quad \text{if } x \neq y, \quad \text{(Mismatch)}$$

$$\delta\begin{pmatrix} x \\ - \end{pmatrix} = \delta\begin{pmatrix} - \\ y \end{pmatrix} = -\rho \quad \text{if } x \neq y \quad \text{(Deletion/Insertion)}.$$

In alignments we don't allow $\begin{pmatrix} - \\ - \end{pmatrix}$.

Example

For $\mu = \rho = 1$ the alignment

$$\begin{array}{cccccccccc} A & T & - & C & - & T & G & A & T \\ - & T & G & C & A & T & - & A & - \end{array}$$

has total score

Example

For $\mu = \rho = 1$ the alignment

$$\begin{array}{cccccccccc} A & T & - & C & - & T & G & A & T \\ - & T & G & C & A & T & - & A & - \end{array}$$

has total score -1.

Problem

For $\mu = \infty, \rho = 0$ find the maximum score of an alignment between the words:

$$\begin{array}{lcccccc} V = & A & T & C & T & G & A & T \\ W = & T & G & C & A & T & A & \end{array}$$

Problem

For $\mu = \infty, \rho = 0$ find the maximum score of an alignment between the words:

$$\begin{array}{lcccccc} V = & A & T & C & T & G & A & T \\ W = & T & G & C & A & T & A & \end{array}$$

Solution

The maximum score for this choice of μ and ρ is equal to the length of a longest common subsequence of V, W, namely 4.

Local sequence alignment problem

Given two words

$$V = v_1 v_2 \ldots v_m,$$

$$W = w_1 w_2 \ldots w_n$$

and constants $\mu, \rho > 0$, find a maximum scoring alignment between *substrings*

$$V' = v_i v_{i+1} \ldots v_{i'},$$

$$W' = w_j w_{j+1} \ldots w_{j'}$$

of V and W.

Local sequence alignment problem

Given two words

$$V = v_1 v_2 \ldots v_m,$$

$$W = w_1 w_2 \ldots w_n$$

and constants $\mu, \rho > 0$, find a maximum scoring alignment between *substrings*

$$V' = v_i v_{i+1} \ldots v_{i'},$$

$$W' = w_j w_{j+1} \ldots w_{j'}$$

of V and W. This maximum score is a measure of similarity $s_{\mu,\rho}(V, W)$ often used in biology.

Local sequence alignment problem

Given two words

$$V = v_1 v_2 \ldots v_m,$$

$$W = w_1 w_2 \ldots w_n$$

and constants $\mu, \rho > 0$, find a maximum scoring alignment between *substrings*

$$V' = v_i v_{i+1} \ldots v_{i'},$$

$$W' = w_j w_{j+1} \ldots w_{j'}$$

of V and W. This maximum score is a measure of similarity $s_{\mu,\rho}(V, W)$ often used in biology.

For $V = TGCATA$, $W = ATCTGAT$ we have a local alignment

$$\begin{array}{ccccc} T & G & C & A & T \\ T & G & - & A & T \end{array}$$

Smith-Waterman Algorithm

Given two words $V = v_1 v_2 ... v_m$, $W = w_1 w_2 ... w_n$ and constants $\mu, \rho > 0$, a maximum scoring local alignment can be found by first constructing an $(m+1) \times (n+1)$ matrix H with entries $H_{i,j}$ defined as follows.

1. $H_{i,0} = 0$ and $H_{0,j} = 0$ for all $0 \leq i \leq m, 0 \leq j \leq n$.
2.

$$H_{i,j} = \max \begin{cases} 0 \\ H_{i-1,j} + \delta(v_i, -) \ \ (\text{Deletion}) \\ H_{i,j-1} + \delta(-, w_j) \ \ (\text{Insertion}) \\ H_{i-1,j-1} + \delta(v_i, w_j) \ \ (\text{Match/Mismatch}) \end{cases}$$

Smith-Waterman Algorithm

Given two words $V = v_1 v_2 ... v_m$, $W = w_1 w_2 ... w_n$ and constants $\mu, \rho > 0$, a maximum scoring local alignment can be found by first constructing an $(m+1) \times (n+1)$ matrix H with entries $H_{i,j}$ defined as follows.

1. $H_{i,0} = 0$ and $H_{0,j} = 0$ for all $0 \leq i \leq m, 0 \leq j \leq n$.
2.

$$H_{i,j} = \max \begin{cases} 0 \\ H_{i-1,j} + \delta(v_i, -) \ \ (\text{Deletion}) \\ H_{i,j-1} + \delta(-, w_j) \ \ (\text{Insertion}) \\ H_{i-1,j-1} + \delta(v_i, w_j) \ \ (\text{Match/Mismatch}) \end{cases}$$

CLAIM: The largest entry in the matrix H is the maximum possible score of a local alignment.

Smith-Waterman Algorithm Illustration

For $V = TGCATA$, $W = ATCTGAT$ and $\mu = \rho = 1$ we get the following 7×8 matrix:

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

Smith-Waterman Algorithm Illustration

For $V = TGCATA$, $W = ATCTGAT$ and $\mu = \rho = 1$ we get the following 7×8 matrix:

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

$$s_{\mu,\rho}(V, W) = 3$$

Similarity vs Dissimilarity

A similarity measure $s(V, W)$ on a finite data set can be normalized to satisfy

$$0 \leq s(V, W) \leq 1 \ .$$

Similarity vs Dissimilarity

A similarity measure $s(V, W)$ on a finite data set can be normalized to satisfy

$$0 \leq s(V, W) \leq 1 \ .$$

We can then define a dissimilarity measure

$$d(V, W) = 1 - s(V, W) \ .$$

Similarity vs Dissimilarity

A similarity measure $s(V, W)$ on a finite data set can be normalized to satisfy

$$0 \leq s(V, W) \leq 1 \ .$$

We can then define a dissimilarity measure

$$d(V, W) = 1 - s(V, W) \ .$$

A dissimilarity measure is a metric if the three metric axioms hold.