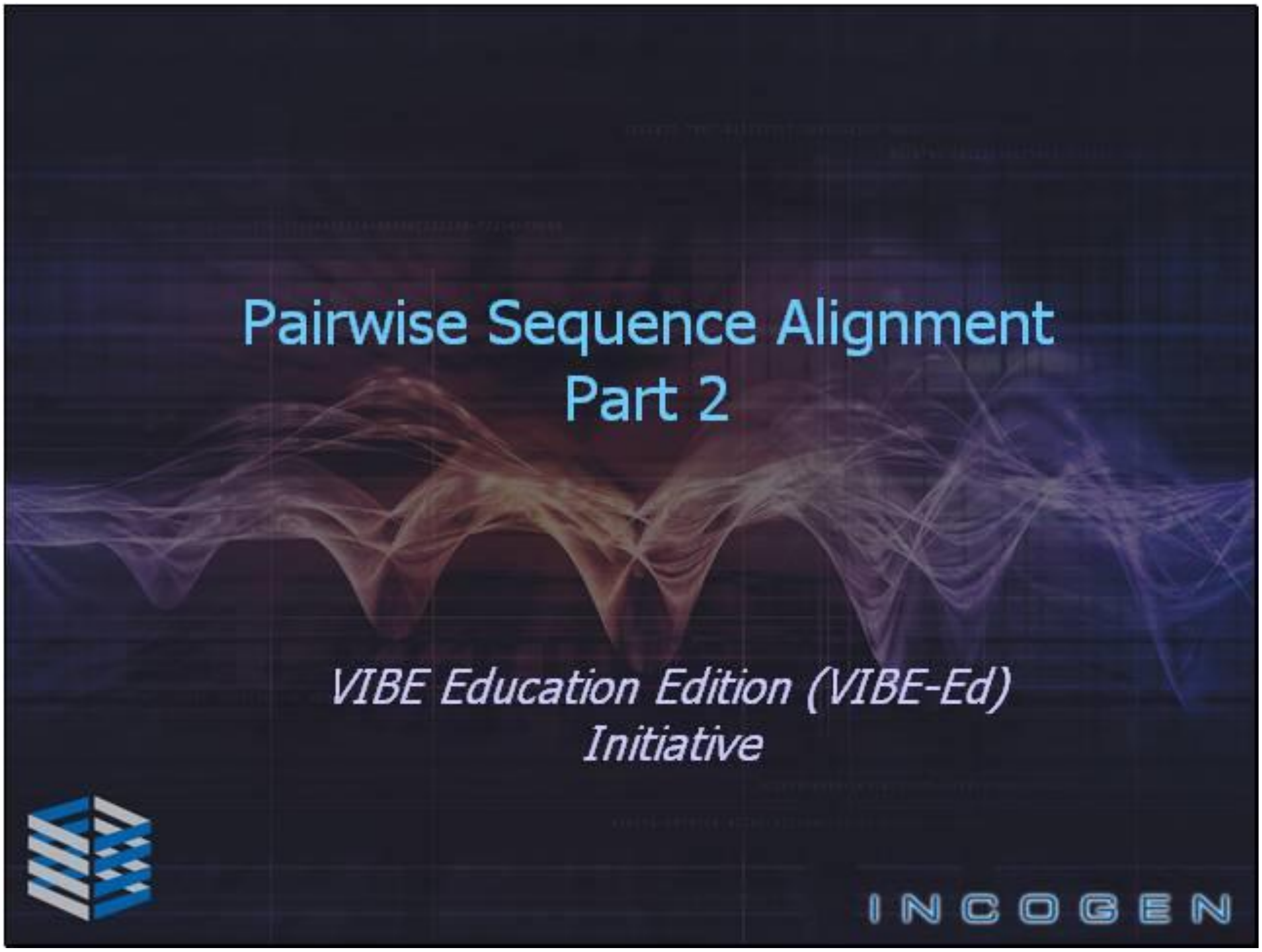
Slide 1 - Pairwise Sequence Alignment Part 2



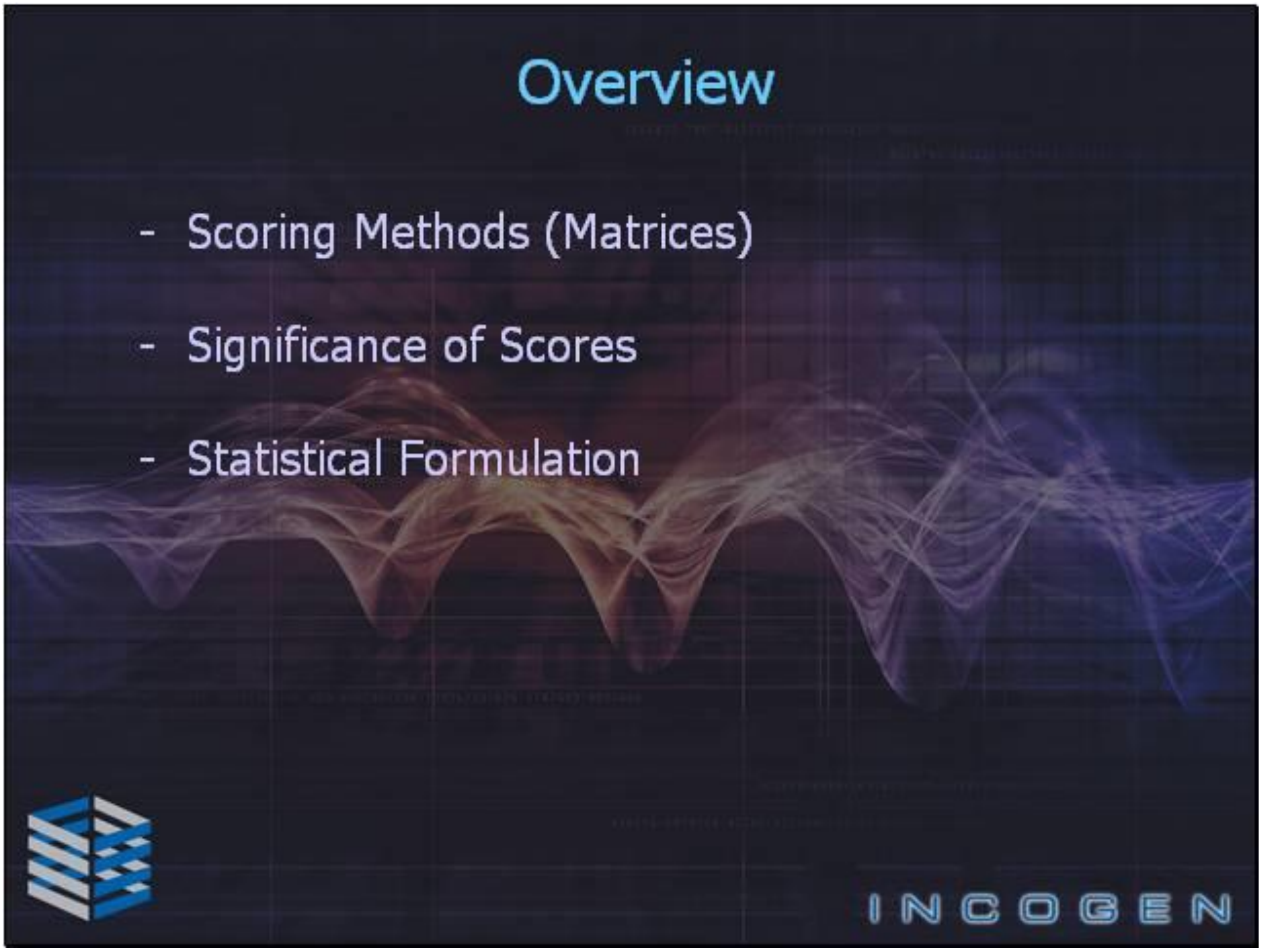
Pairwise Sequence Alignment
Part 2

*VIBE Education Edition (VIBE-Ed)
Initiative*

Slide notes

This presentation continues the discussion of pairwise sequence alignment.

Slide 2 - Overview



Overview

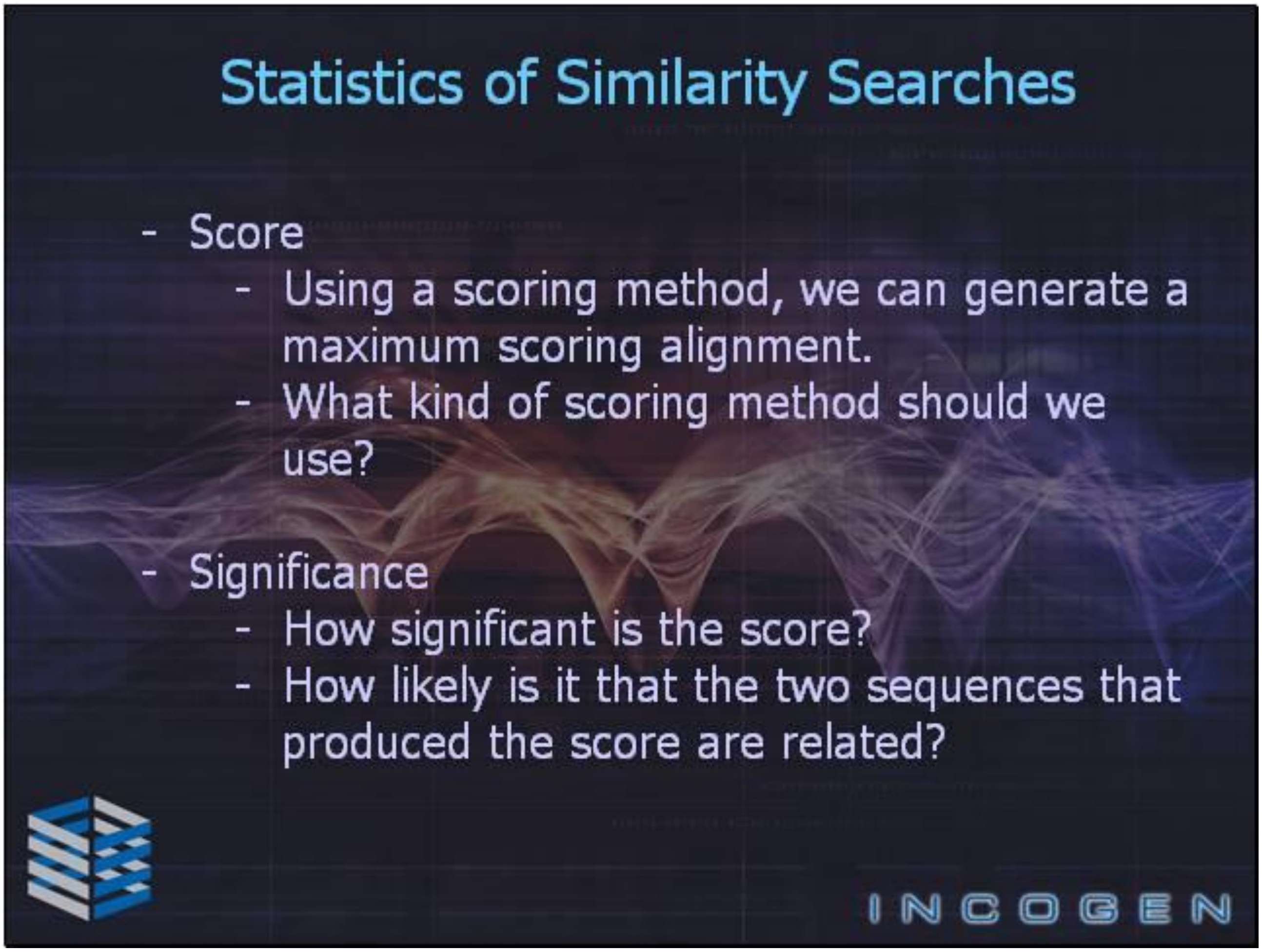
- Scoring Methods (Matrices)
- Significance of Scores
- Statistical Formulation

INCOGEN

Slide notes

In particular, we are going to have a more in-depth discussion about scoring methods, about the significance of the scores that the algorithms produce, and about the statistics behind the calculations of significance and score.

Slide 3 - Statistics of Similarity Searches



Statistics of Similarity Searches

- Score
  - Using a scoring method, we can generate a maximum scoring alignment.
  - What kind of scoring method should we use?
- Significance
  - How significant is the score?
  - How likely is it that the two sequences that produced the score are related?

Slide notes

Specifically, we want to think about the following questions. First, what kind of scoring method should we use to generate the maximum scoring alignment? Second, how significant is this score; that is, how likely is it that the two sequences producing this alignment score are related?

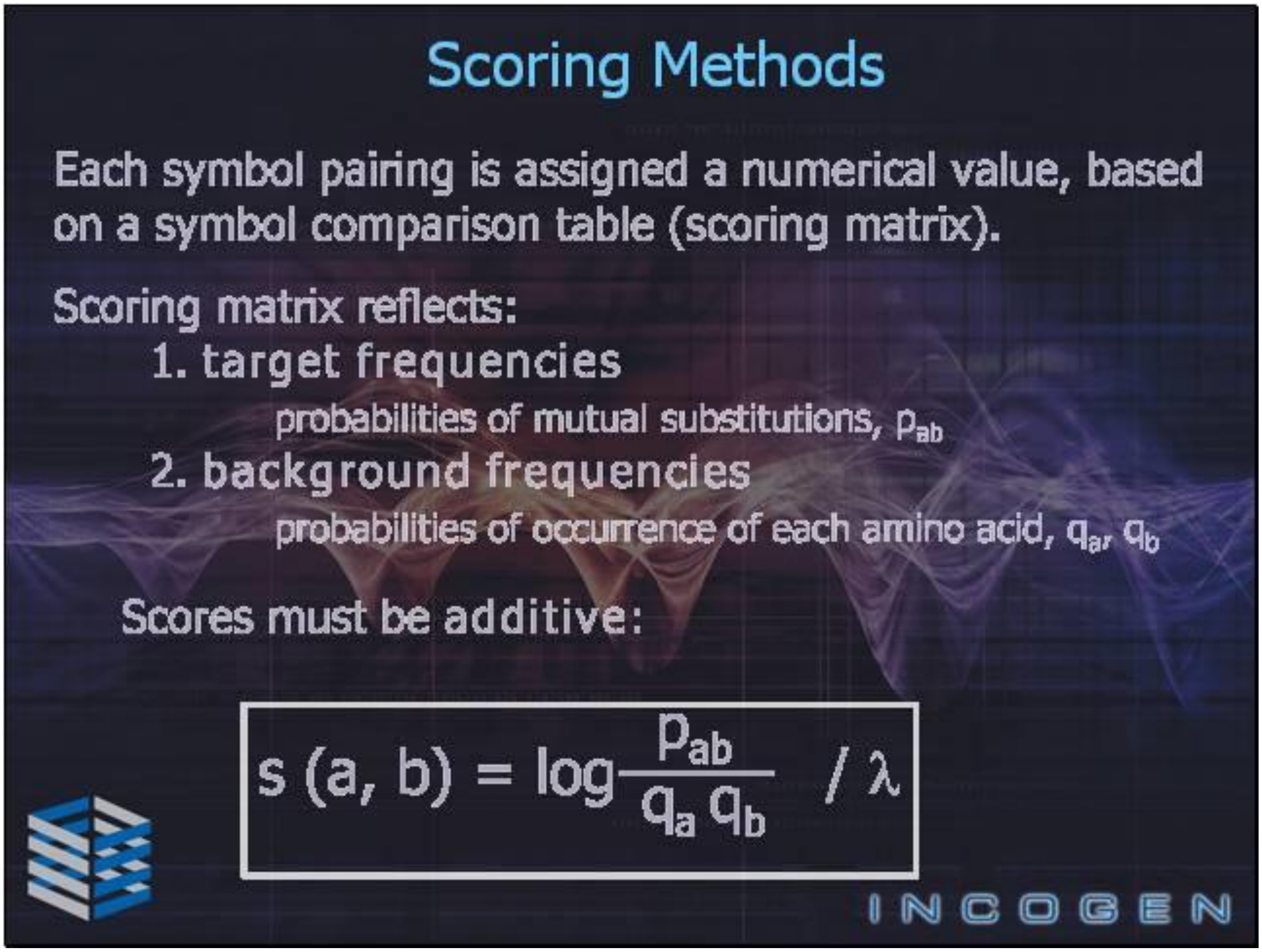
Slide 4 - Scoring Methods

Scoring Methods

Slide notes

First, some more detailed information about scoring methods.

Slide 5 - Slide 5



Scoring Methods

Each symbol pairing is assigned a numerical value, based on a symbol comparison table (scoring matrix).

Scoring matrix reflects:

1. target frequencies

   probabilities of mutual substitutions, p_{ab}

2. background frequencies

   probabilities of occurrence of each amino acid, q_a, q_b

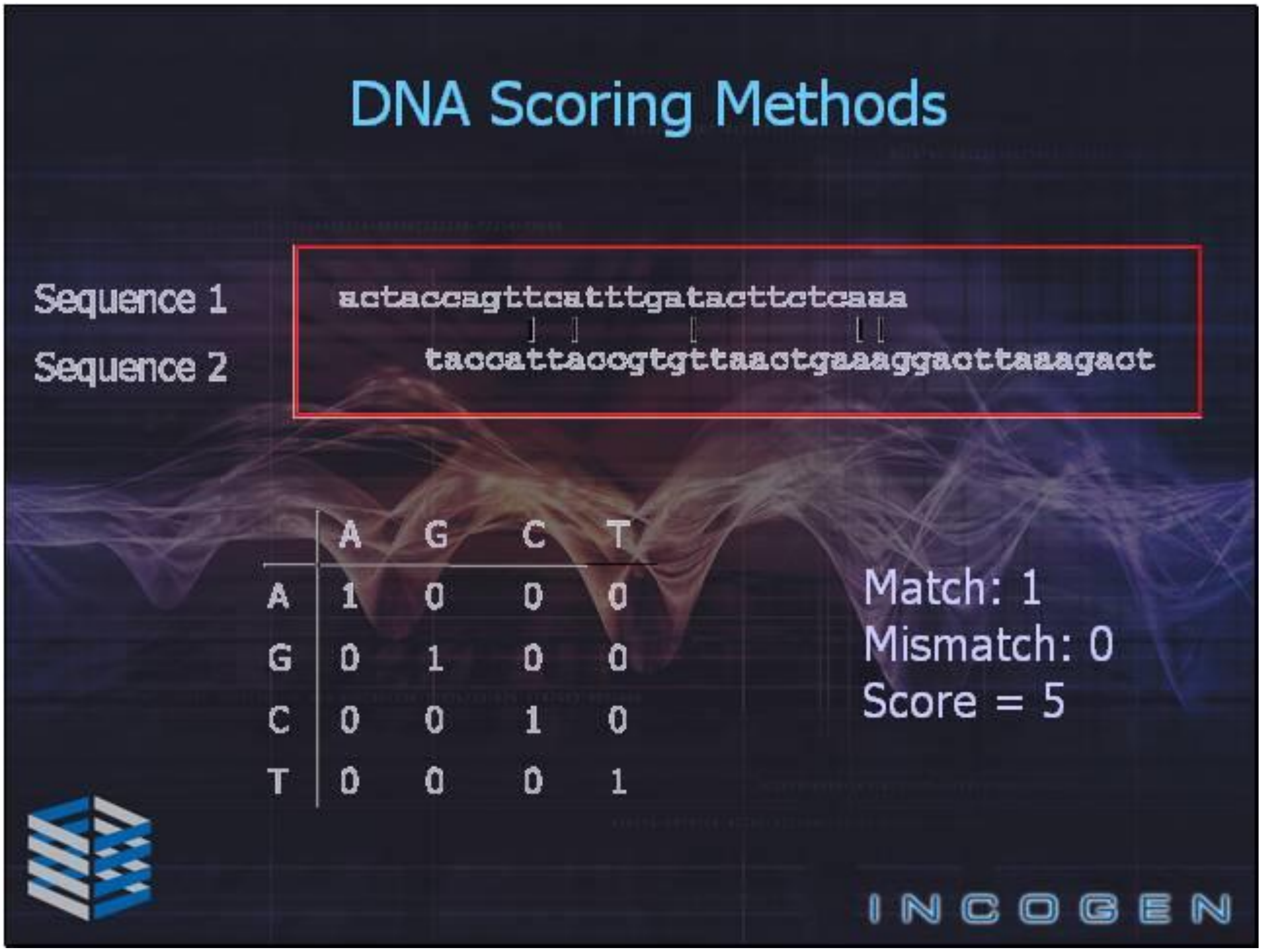
Scores must be additive:

$$\boxed{s\,(a, b) = \log \frac{p_{ab}}{q_a\, q_b} \;\; / \, \lambda}$$

Slide notes

To calculate the final score of a potential alignment, each symbol pairing is assigned a numerical value, based on how frequently you would expect to see this pairing in two related species. To simplify calculating the final score, the values must be additive. In general, the values are arranged in a “scoring matrix.”

Slide 6 - Slide 6



DNA Scoring Methods

```
Sequence 1   actaccagttcatttgatacttctcaaa
                      |  |     |       ||
Sequence 2       taccattaccgtgttaactgaaaggacttaaagact
```

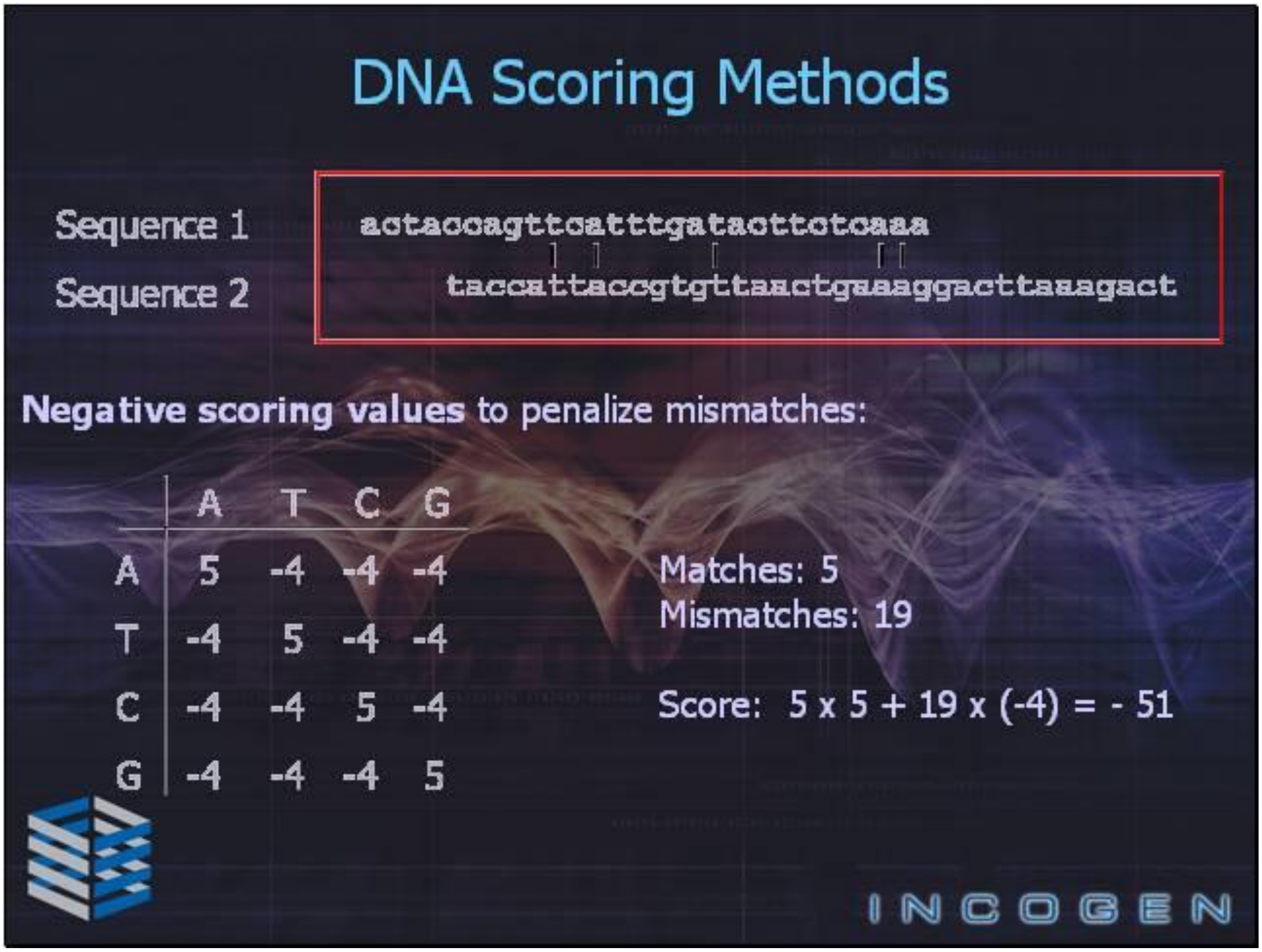
$$\begin{array}{c|cccc} & A & G & C & T \\ \hline A & 1 & 0 & 0 & 0 \\ G & 0 & 1 & 0 & 0 \\ C & 0 & 0 & 1 & 0 \\ T & 0 & 0 & 0 & 1 \end{array}$$

Match: 1
Mismatch: 0
Score = 5

Slide notes

The simplest scoring matrix for DNA sequences assigns a value of 1 for a match and 0 for a mismatch. Using this matrix, the example alignment shown has a score of 5.

Slide 7 - Slide 7



DNA Scoring Methods

Sequence 1

Sequence 2

```
actaccagttcatttgatacttctcaaa
         | |     |       ||
    taccattaccgtgttaactgaaaggacttaaagact
```

Negative scoring values to penalize mismatches:

	A	T	C	G
A	5	-4	-4	-4
T	-4	5	-4	-4
C	-4	-4	5	-4
G	-4	-4	-4	5

Matches: 5
Mismatches: 19

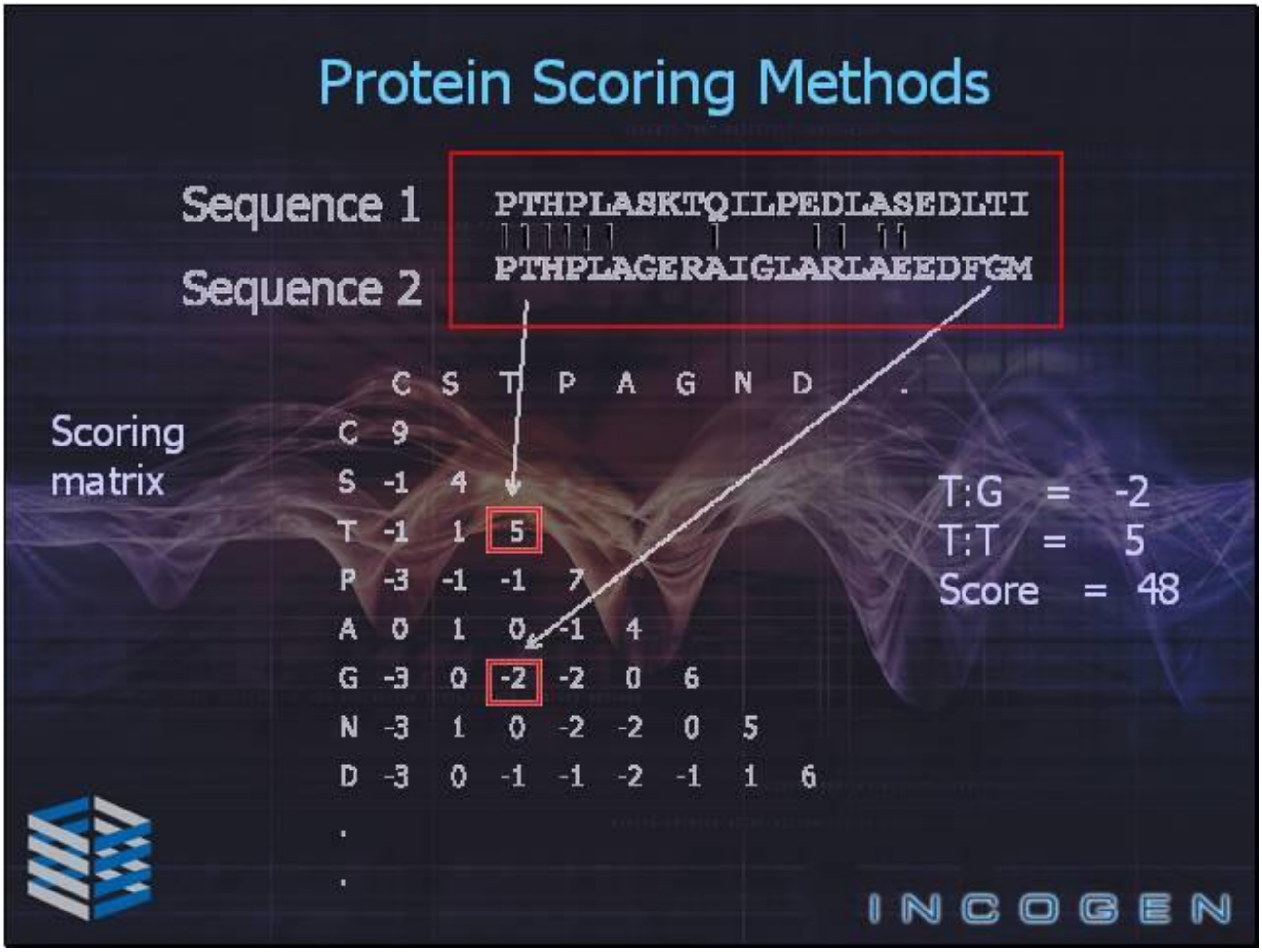
Score: 5 x 5 + 19 x (-4) = - 51

INCOGEN

Slide notes

Using this scoring matrix makes it more difficult to tell the difference between a very good alignment that is short and a longer alignment that is very poor. So, we can introduce negative scoring values to penalize mismatches. For this scoring matrix, a match gets a score of 5, and a mismatch gets a score of -4. Using this modified matrix, the example alignment has a score of -51.

Slide 8 - Slide 8



Protein Scoring Methods

Sequence 1 PTHPLASKTQILPEDLASEDLTI

Sequence 2 PTHPLAGERAIGLARLAEEDFGM

Scoring matrix

	C	S	T	P	A	G	N	D	.
C	9								
S	-1	4							
T	-1	1	5						
P	-3	-1	-1	7					
A	0	1	0	-1	4				
G	-3	0	-2	-2	0	6			
N	-3	1	0	-2	-2	0	5		
D	-3	0	-1	-1	-2	-1	1	6	
.									
.									

T:G = -2

T:T = 5

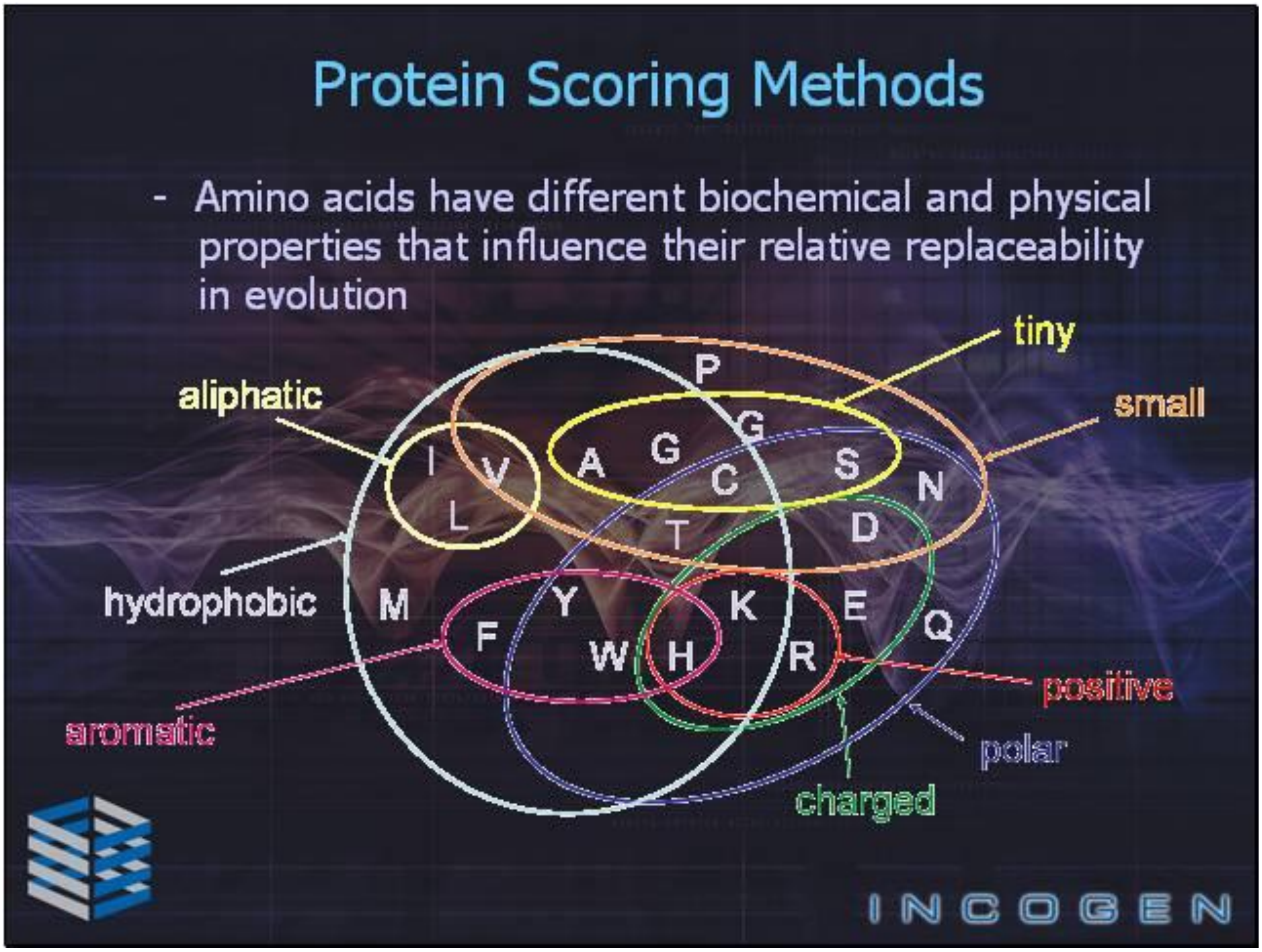
Score = 48

INCOGEN

Slide notes

Protein scoring matrices are larger and more complicated looking since there is a different probability for each amino acid pairing. To find the value associated with any amino acid substitution pair, find the two amino acids on the edges of the table and read off the value where the row and column intersect. For this example scoring matrix, let's look at Threonine, whose single-letter code is T, and Glycine, whose single-letter code is G. The value for a T:T match is 5, and the value for a T:G mismatch is -2.

Slide 9 - Slide 9



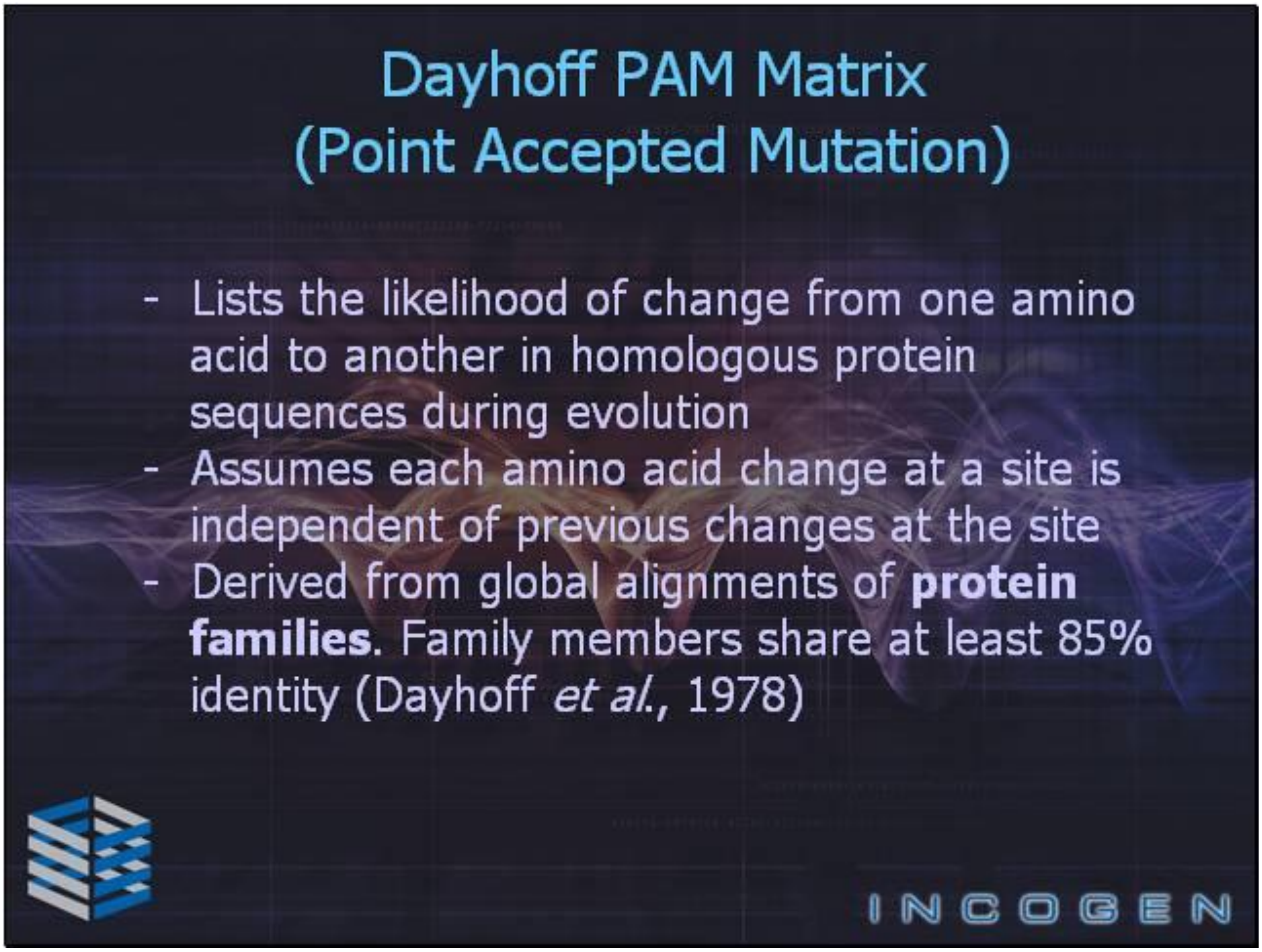
Protein Scoring Methods

- Amino acids have different biochemical and physical properties that influence their relative replaceability in evolution

Slide notes

Different amino acid pairings have different values because amino acids have different biochemical and physical properties that influence their relative substitution rate in evolution. In this image, the amino acids are grouped according to the chemistry of their side groups.

Slide 10 - Dayhoff PAM Matrix (Point Accepted Mutation)



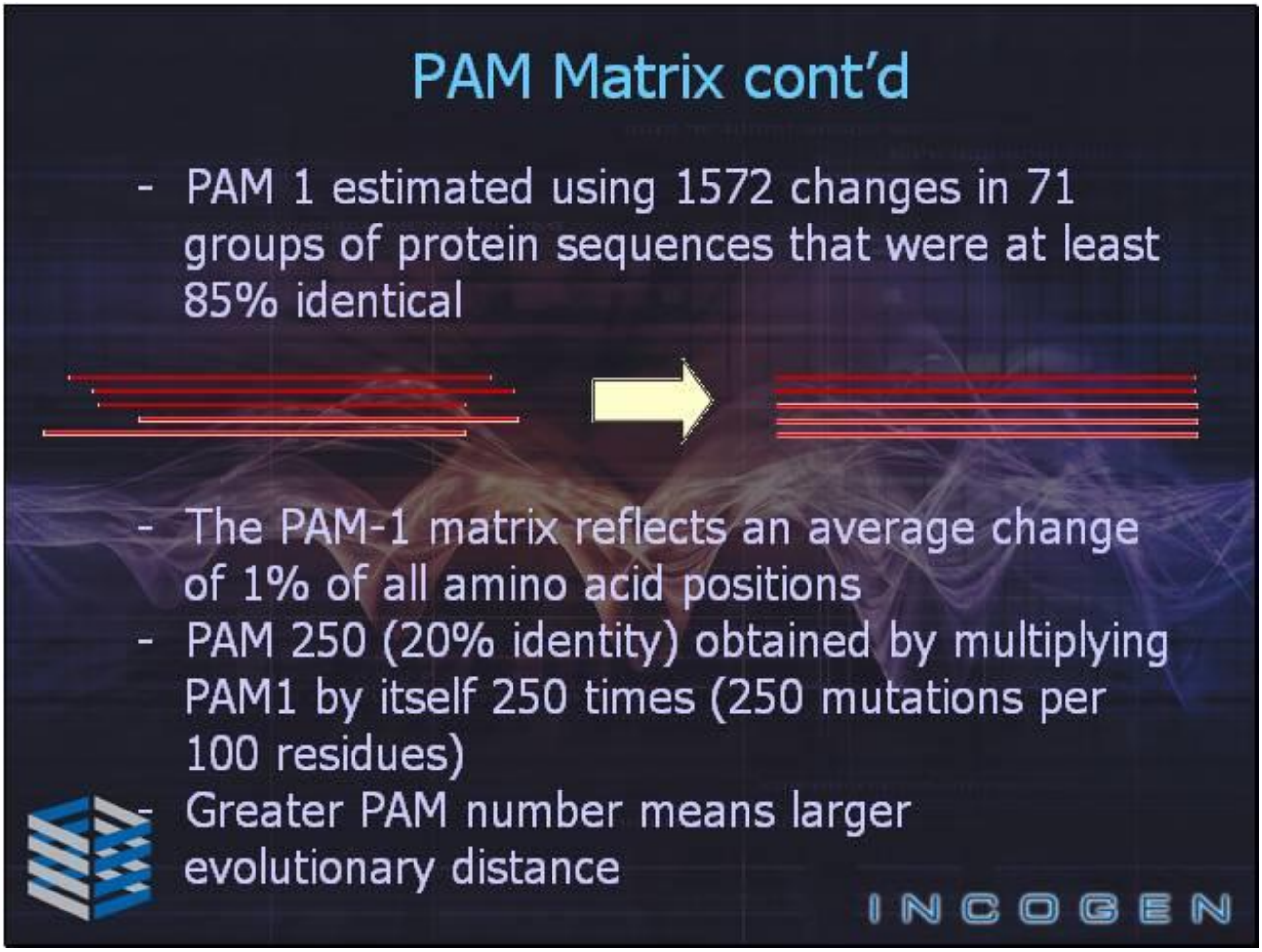
Dayhoff PAM Matrix (Point Accepted Mutation)

- Lists the likelihood of change from one amino acid to another in homologous protein sequences during evolution
- Assumes each amino acid change at a site is independent of previous changes at the site
- Derived from global alignments of **protein families**. Family members share at least 85% identity (Dayhoff *et al.*, 1978)

Slide notes

There has been a lot of work to develop meaningful score matrices for protein sequence alignment. The first family of scoring matrices we will look at is the Dayhoff PAM matrix. The matrices list the likelihood that one amino acid could change to another in homologous protein sequences during evolution. It assumes that each amino acid change is independent of any previous changes that occurred at that site. This matrix family was derived from global alignments of protein families that are at least 85% identical.

Slide 11 - PAM Matrix cont’d



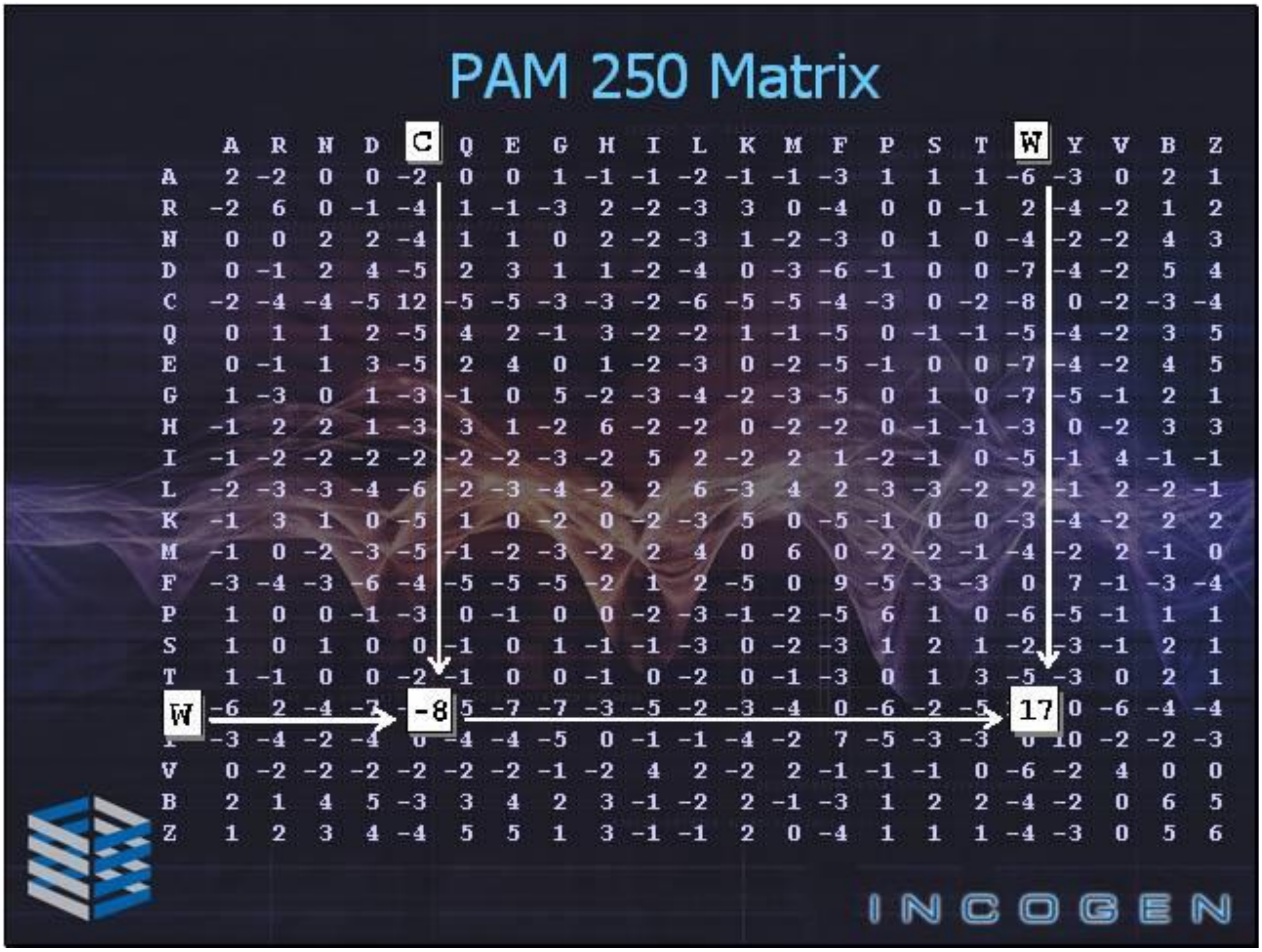
PAM Matrix cont’d

- PAM 1 estimated using 1572 changes in 71 groups of protein sequences that were at least 85% identical
- The PAM-1 matrix reflects an average change of 1% of all amino acid positions
- PAM 250 (20% identity) obtained by multiplying PAM1 by itself 250 times (250 mutations per 100 residues)
- Greater PAM number means larger evolutionary distance

Slide notes

This first matrix, which represents 85% identity, is called PAM 1. PAM 1 reflects an average change of 1% of all amino acid positions. Since we are generally looking at alignments between sequences that are less than 85% similar, other PAM matrices, such as PAM 250, were calculated by multiplying PAM 1 by itself for some number of times; the number of times is reflected in the name of the matrix. Larger PAM numbers mean that that the matrix represents a larger evolutionary distance.

Slide 12 - Slide 12



PAM 250 Matrix

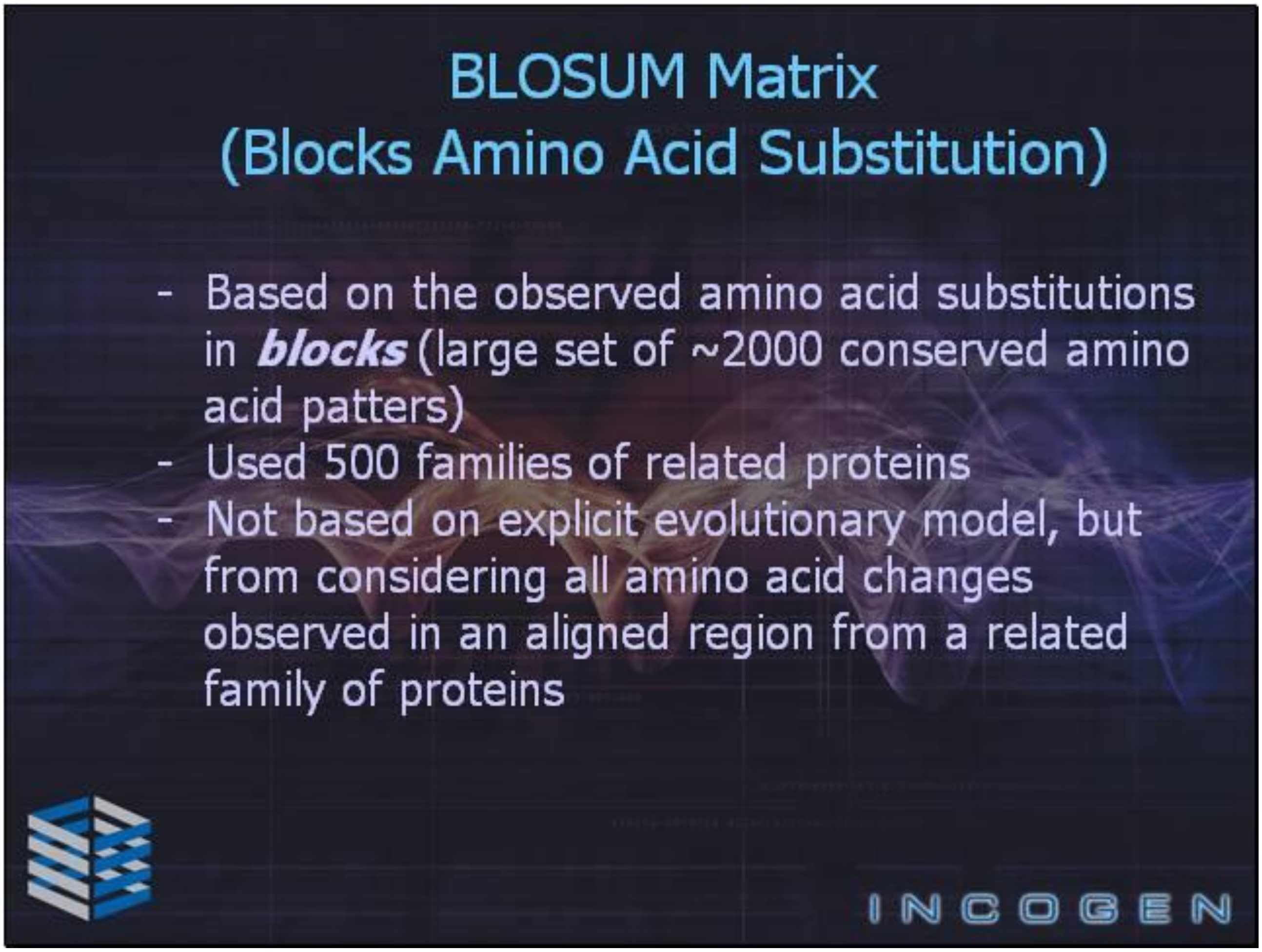
	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z
A	2	-2	0	0	-2	0	0	1	-1	-1	-2	-1	-1	-3	1	1	1	-6	-3	0	2	1
R	-2	6	0	-1	-4	1	-1	-3	2	-2	-3	3	0	-4	0	0	-1	2	-4	-2	1	2
N	0	0	2	2	-4	1	1	0	2	-2	-3	1	-2	-3	0	1	0	-4	-2	-2	4	3
D	0	-1	2	4	-5	2	3	1	1	-2	-4	0	-3	-6	-1	0	0	-7	-4	-2	5	4
C	-2	-4	-4	-5	12	-5	-5	-3	-3	-2	-6	-5	-5	-4	-3	0	-2	-8	0	-2	-3	-4
Q	0	1	1	2	-5	4	2	-1	3	-2	-2	1	-1	-5	0	-1	-1	-5	-4	-2	3	5
E	0	-1	1	3	-5	2	4	0	1	-2	-3	0	-2	-5	-1	0	0	-7	-4	-2	4	5
G	1	-3	0	1	-3	-1	0	5	-2	-3	-4	-2	-3	-5	0	1	0	-7	-5	-1	2	1
H	-1	2	2	1	-3	3	1	-2	6	-2	-2	0	-2	-2	0	-1	-1	-3	0	-2	3	3
I	-1	-2	-2	-2	-2	-2	-2	-3	-2	5	2	-2	2	1	-2	-1	0	-5	-1	4	-1	-1
L	-2	-3	-3	-4	-6	-2	-3	-4	-2	2	6	-3	4	2	-3	-3	-2	-2	-1	2	-2	-1
K	-1	3	1	0	-5	1	0	-2	0	-2	-3	5	0	-5	-1	0	0	-3	-4	-2	2	2
M	-1	0	-2	-3	-5	-1	-2	-3	-2	2	4	0	6	0	-2	-2	-1	-4	-2	2	-1	0
F	-3	-4	-3	-6	-4	-5	-5	-5	-2	1	2	-5	0	9	-5	-3	-3	0	7	-1	-3	-4
P	1	0	0	-1	-3	0	-1	0	0	-2	-3	-1	-2	-5	6	1	0	-6	-5	-1	1	1
S	1	0	1	0	0	-1	0	1	-1	-1	-3	0	-2	-3	1	2	1	-2	-3	-1	2	1
T	1	-1	0	0	-2	-1	0	0	-1	0	-2	0	-1	-3	0	1	3	-5	-3	0	2	1
W	-6	2	-4	-7	-8	-5	-7	-7	-3	-5	-2	-3	-4	0	-6	-2	-5	17	0	-6	-4	-4
Y	-3	-4	-2	-4	0	-4	-4	-5	0	-1	-1	-4	-2	7	-5	-3	-3	0	10	-2	-2	-3
V	0	-2	-2	-2	-2	-2	-2	-1	-2	4	2	-2	2	-1	-1	-1	0	-6	-2	4	0	0
B	2	1	4	5	-3	3	4	2	3	-1	-2	2	-1	-3	1	2	2	-4	-2	0	6	5
Z	1	2	3	4	-4	5	5	1	3	-1	-1	2	0	-4	1	1	1	-4	-3	0	5	6

Slide notes

Here is the PAM 250 matrix. This matrix is symmetric, meaning that the value in the matrix is independent of which amino acid out of a pair substituted for the other.

Let's look at Tryptophan, represented by its single-letter code, W, and Cysteine, represented by its single-letter code, C. The value associated with a W:W pairing, 17, is a very large positive number, while almost all of the other pairings with W, such as W:C, have relatively large negative scores. Therefore, the matrix represents the fact that Tryptophan is a conserved residue.

Slide 13 - BLOSUM Matrix (Blocks Amino Acid Substitution)



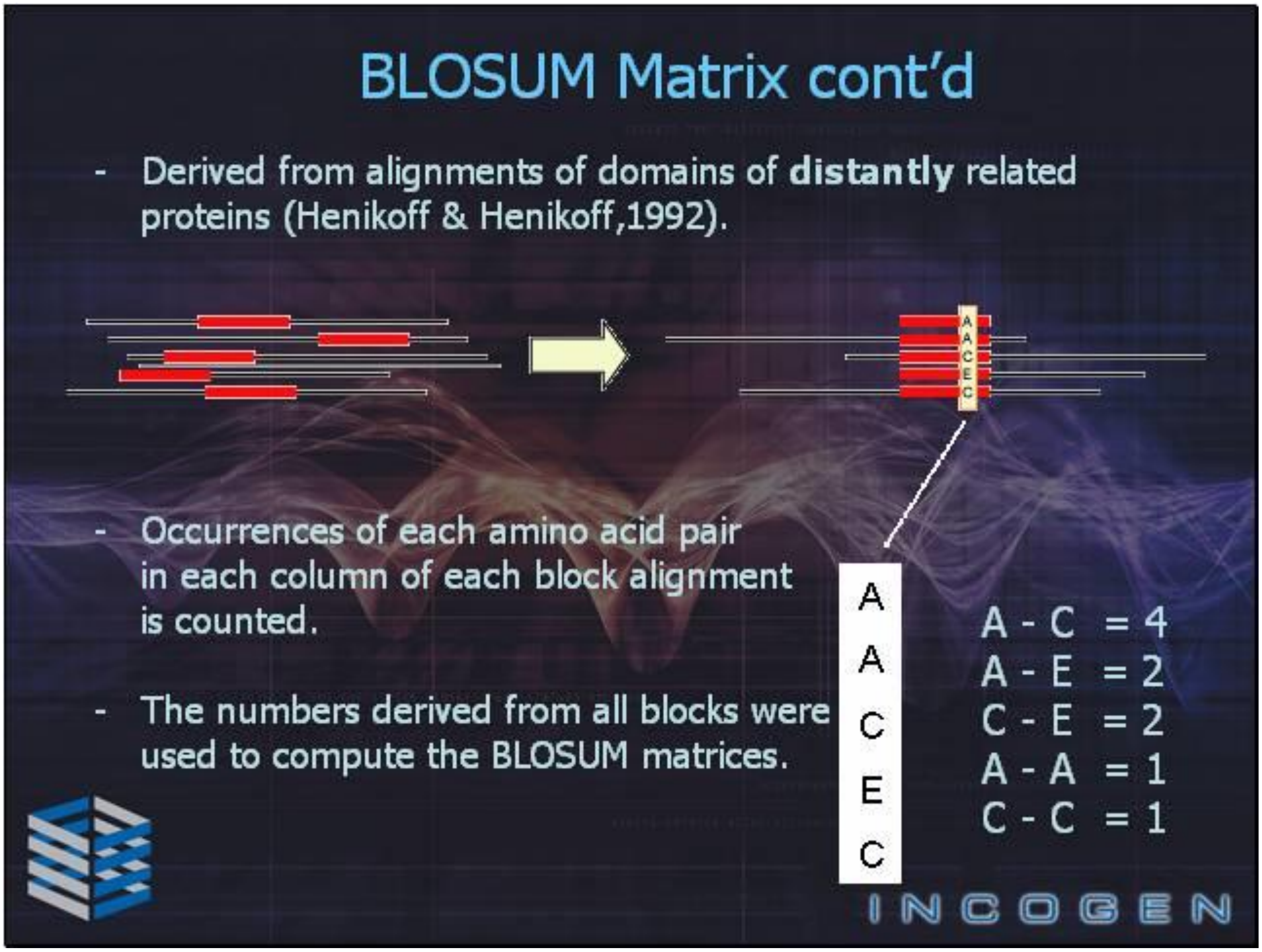
BLOSUM Matrix
(Blocks Amino Acid Substitution)

- Based on the observed amino acid substitutions in ***blocks*** (large set of ~2000 conserved amino acid patters)
- Used 500 families of related proteins
- Not based on explicit evolutionary model, but from considering all amino acid changes observed in an aligned region from a related family of proteins

Slide notes

The second family of scoring matrices is the BLOSUM matrix. The BLOSUM matrices were based on the amino acid substitution rate in highly conserved blocks. The BLOSUM matrices are not explicitly based on an evolutionary model, as the PAM matrices are, but are based instead on related families of proteins.

Slide 14 - Slide 14



BLOSUM Matrix cont'd

- Derived from alignments of domains of **distantly** related proteins (Henikoff & Henikoff,1992).

- Occurrences of each amino acid pair in each column of each block alignment is counted.

- The numbers derived from all blocks were used to compute the BLOSUM matrices.

A
A
C
E
C

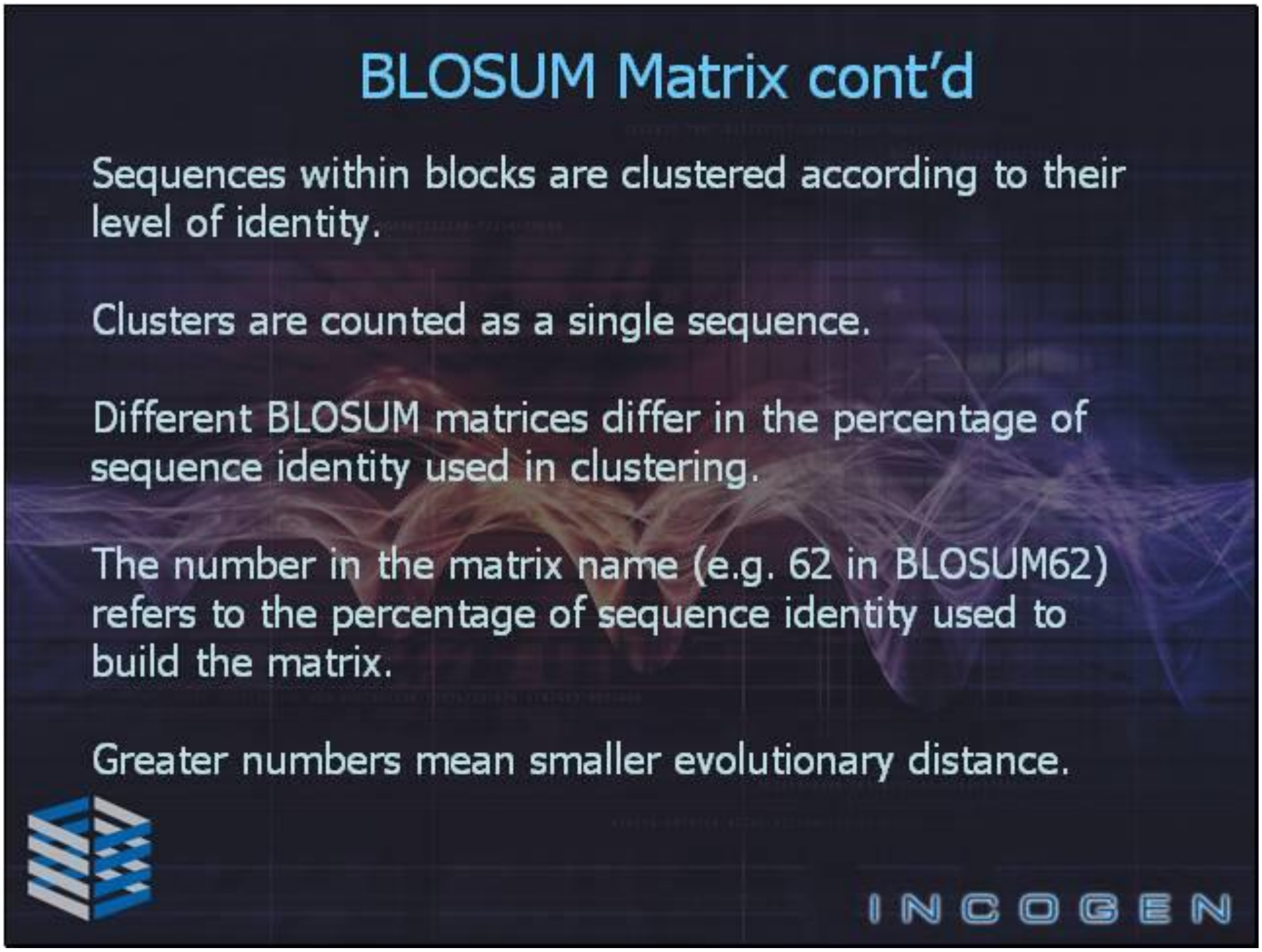
A - C = 4
A - E = 2
C - E = 2
A - A = 1
C - C = 1

INCOGEN

Slide notes

Specifically, the BLOSUM matrices were derived from alignments of domains in distantly related proteins. Occurrences of each amino acid pair in each column of each block were counted, and the totals were used to compute the BLOSUM matrices.

Slide 15 - Slide 15



BLOSUM Matrix cont'd

Sequences within blocks are clustered according to their level of identity.

Clusters are counted as a single sequence.

Different BLOSUM matrices differ in the percentage of sequence identity used in clustering.

The number in the matrix name (e.g. 62 in BLOSUM62) refers to the percentage of sequence identity used to build the matrix.

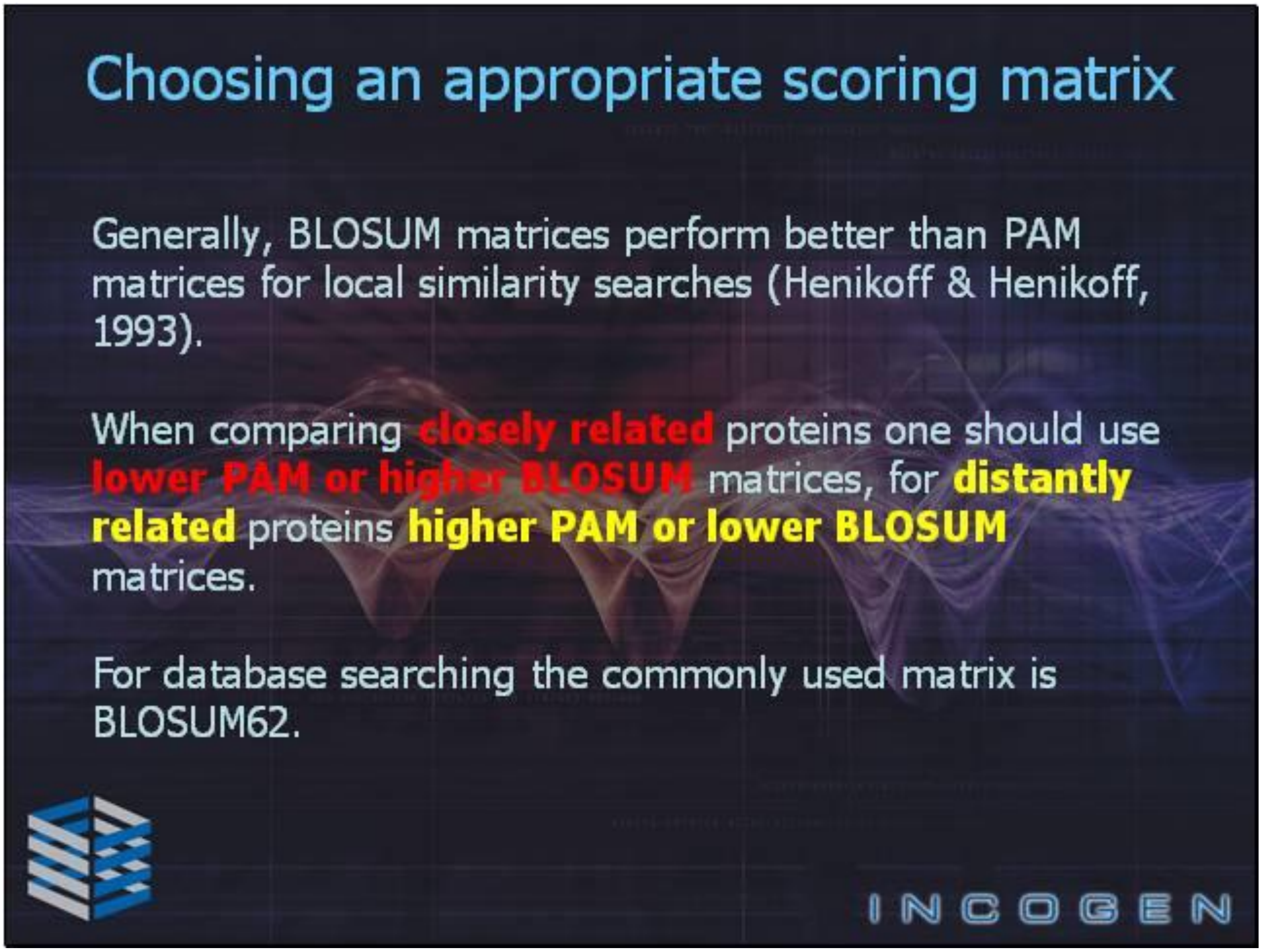
Greater numbers mean smaller evolutionary distance.

INCOGEN

Slide notes

The sequences within blocks were clustered based on their level of identity, and the different BLOSUM matrices differ in the percentage of sequence identity used in the clustering. This percentage is reflected in the matrix name. Unlike the PAM matrices, larger numbers mean a smaller evolutionary distance.

Slide 16 - Slide 16



Choosing an appropriate scoring matrix

Generally, BLOSUM matrices perform better than PAM matrices for local similarity searches (Henikoff & Henikoff, 1993).

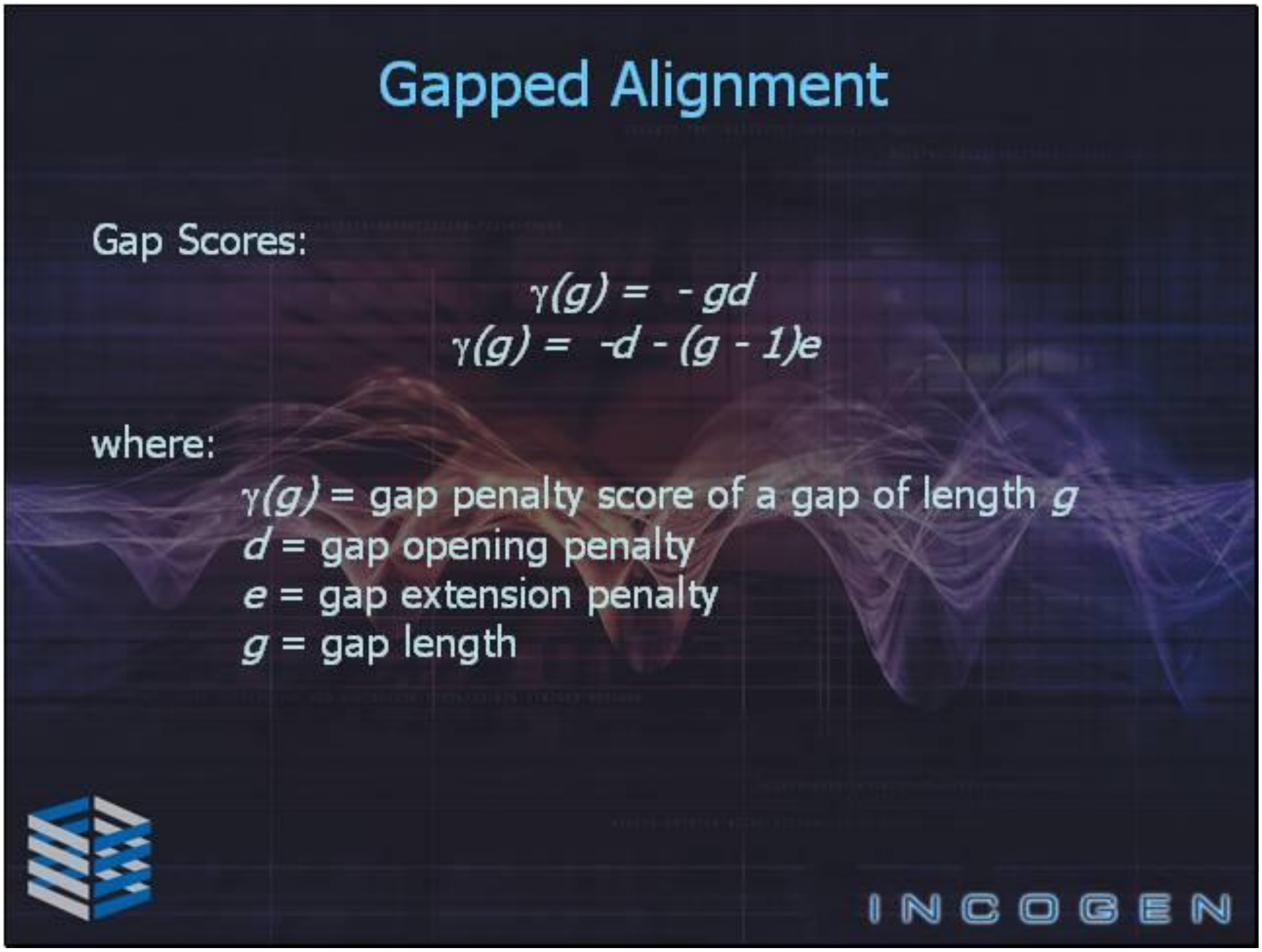
When comparing **closely related** proteins one should use **lower PAM or higher BLOSUM** matrices, for **distantly related** proteins **higher PAM or lower BLOSUM** matrices.

For database searching the commonly used matrix is BLOSUM62.

Slide notes

So, which matrix do we choose for our alignment? Generally, BLOSUM matrices perform better than PAM matrices for local similarity searches. It is very important that you remember that large PAM numbers or small BLOSUM numbers should be used for distantly related proteins, while small PAM numbers or large BLOSUM numbers should be used for closely related proteins. The most commonly used matrix for calculating local alignments is BLOSUM 62.

Slide 17 - Gapped Alignment



Gapped Alignment

Gap Scores:

$$\gamma(g) = -gd$$

$$\gamma(g) = -d - (g - 1)e$$

where:

$\gamma(g)$ = gap penalty score of a gap of length g
d = gap opening penalty
e = gap extension penalty
g = gap length

Slide notes

Gaps are scored independently of the scoring matrix used. There are two methods of scoring gaps. The first simply multiplies the length of the gap by a penalty per length. The second method uses a gap opening penalty added to a gap extension penalty that is multiplied by the amount the gap is extended (that is, the length of the gap minus 1).

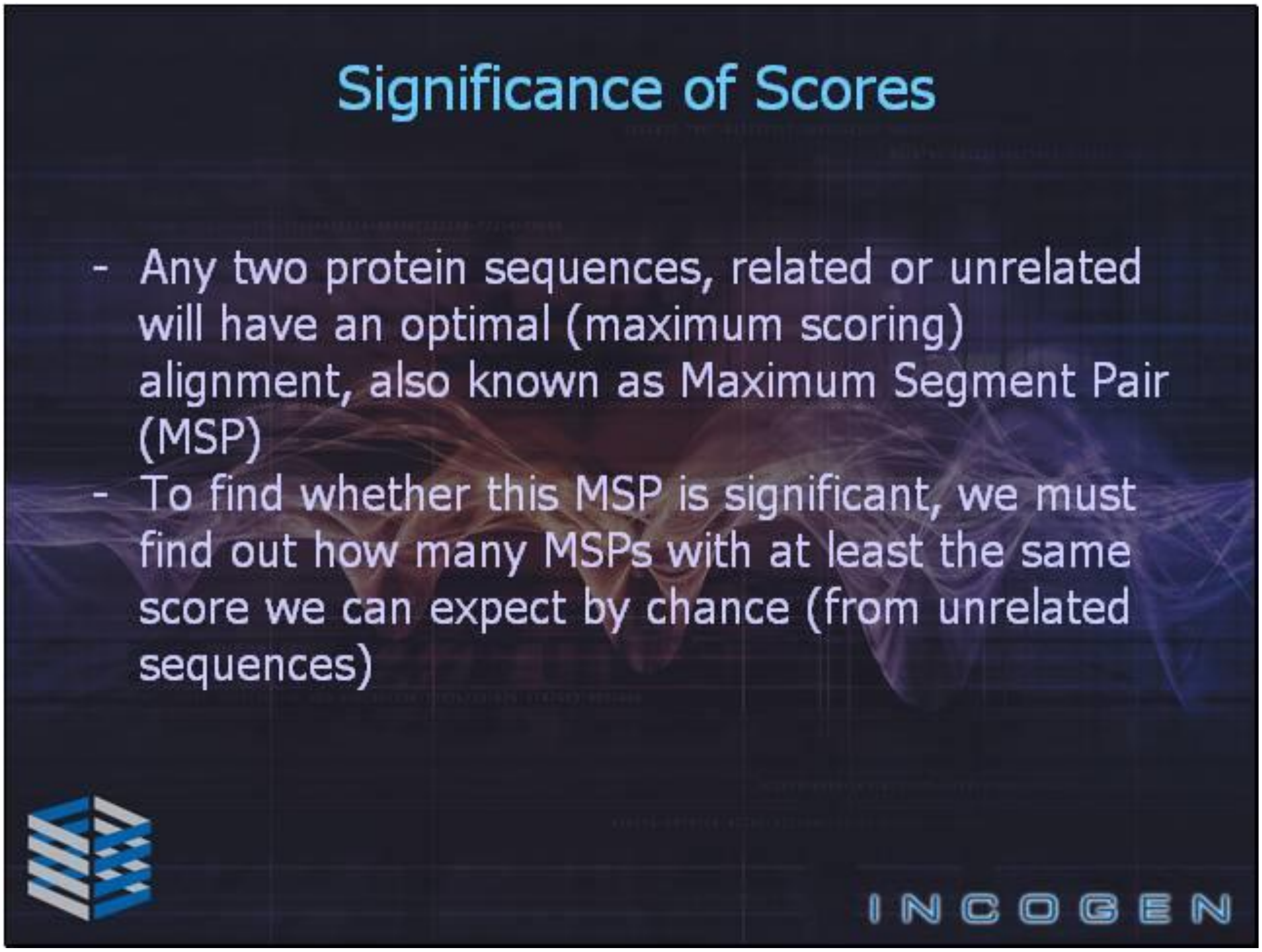
Slide 18 - Significance of Scores

Significance of Scores

Slide notes

Now that we know how scores are calculated, let’s take a brief look at their significance.

Slide 19 - Significance of Scores, con't



Significance of Scores

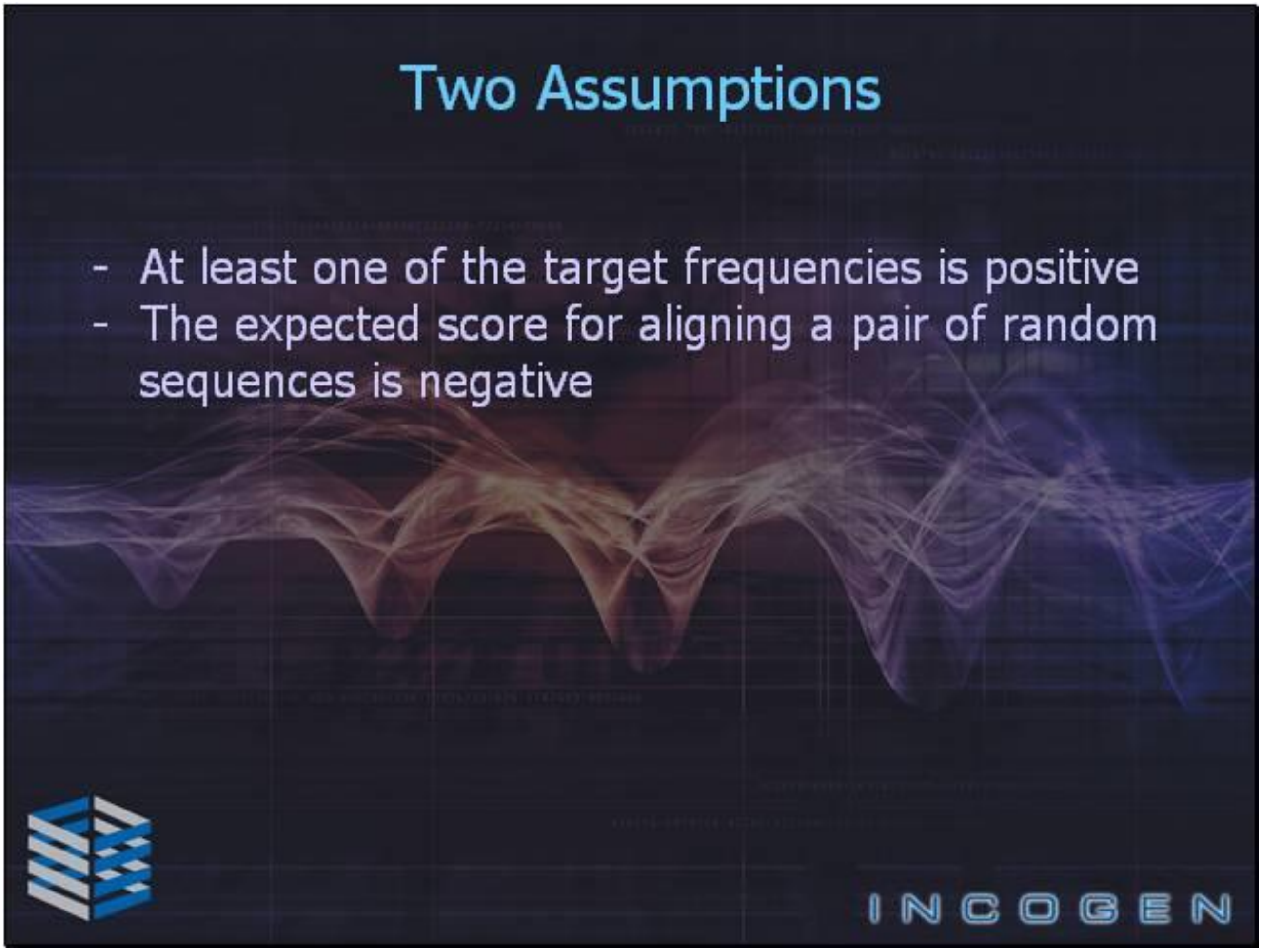
- Any two protein sequences, related or unrelated will have an optimal (maximum scoring) alignment, also known as Maximum Segment Pair (MSP)
- To find whether this MSP is significant, we must find out how many MSPs with at least the same score we can expect by chance (from unrelated sequences)

INCOGEN

Slide notes

It is always possible to find an optimal alignment, or Maximum Segment Pair (MSP), between any two protein sequences, regardless of whether they are related. However, not all MSPs are significant. To determine if the MSP is significant, we need to find out how many MSPs with at least that same score we can expect by chance.

Slide 20 - Two Assumptions



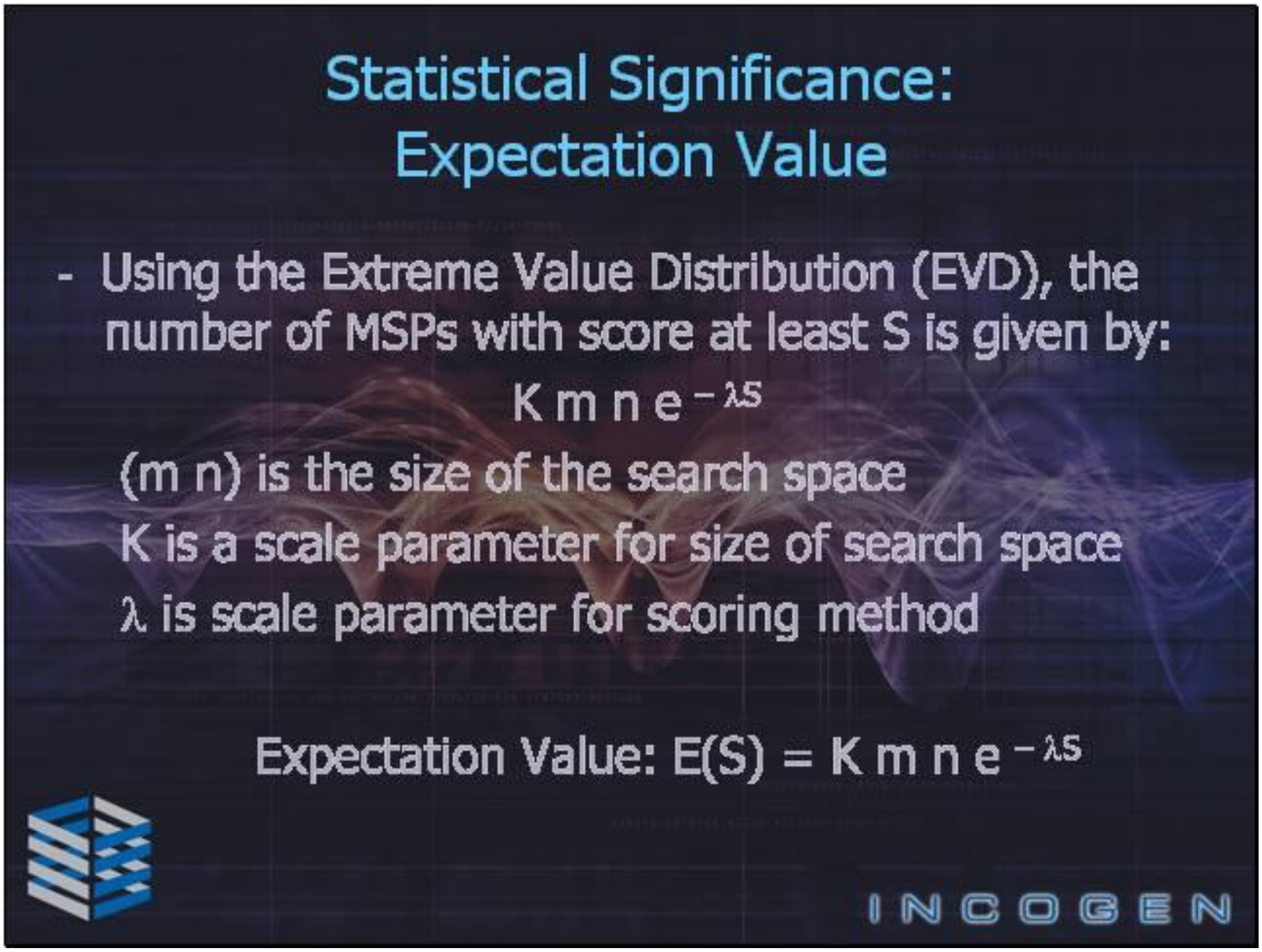
Two Assumptions

- At least one of the target frequencies is positive
- The expected score for aligning a pair of random sequences is negative

Slide notes

To calculate this, we will make two assumptions. First, at least one of the target frequencies is positive. Second, the expected score for aligning a pair of random sequences is negative.

Slide 21 - Statistical Significance: Expectation Value



Statistical Significance: Expectation Value

- Using the Extreme Value Distribution (EVD), the number of MSPs with score at least S is given by:

$$K\, m\, n\, e^{-\lambda S}$$

(m n) is the size of the search space

K is a scale parameter for size of search space

λ is scale parameter for scoring method

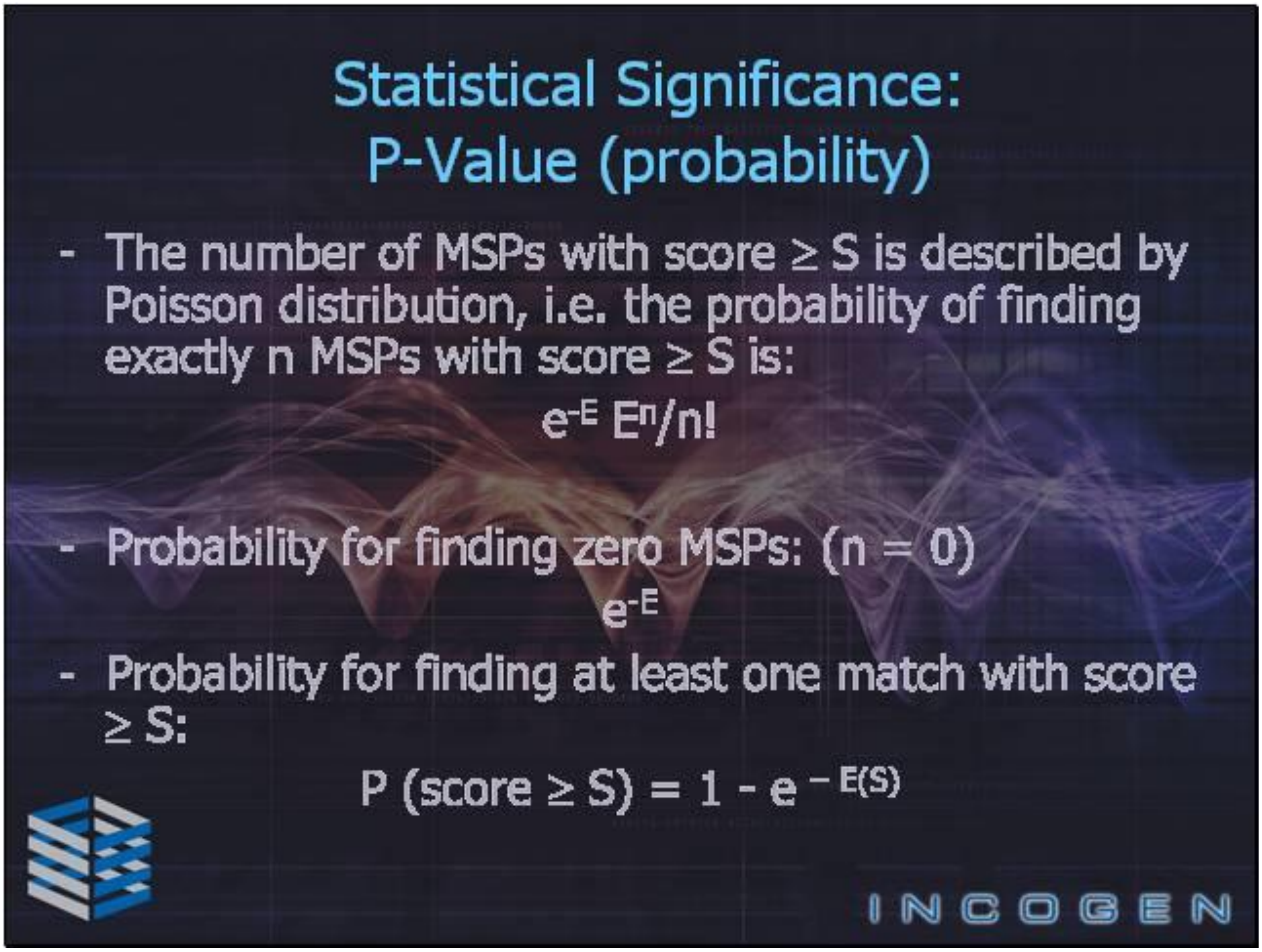
Expectation Value: $E(S) = K\, m\, n\, e^{-\lambda S}$

INCOGEN

Slide notes

With these assumptions, we can use the extreme value distribution (EVD) to calculate the number of alignments between random sequences that we expect to see with a given score or better. We define this quantity to be the e-value. In general, as the score increases, the e-value decreases, indicating that this alignment is more likely to be biologically significant.

Slide 22 - Statistical Significance: P-Value (probability)



Statistical Significance: P-Value (probability)

- The number of MSPs with score ≥ S is described by Poisson distribution, i.e. the probability of finding exactly n MSPs with score ≥ S is:

$$e^{-E} E^{n}/n!$$

- Probability for finding zero MSPs: (n = 0)

$$e^{-E}$$

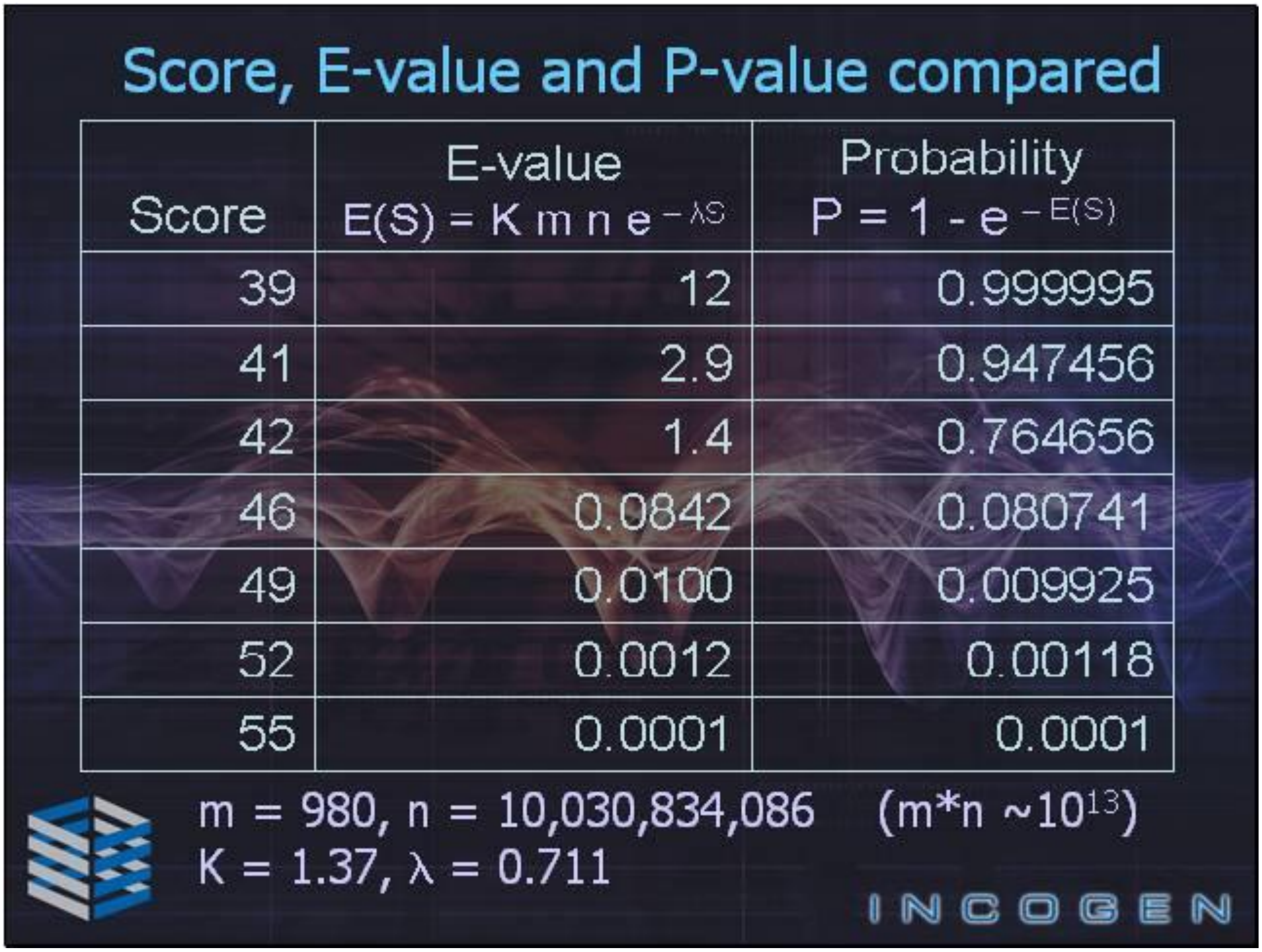
- Probability for finding at least one match with score ≥ S:

$$P\ (\text{score} \geq S) = 1 - e^{-E(S)}$$

Slide notes

We can also calculate the probability of finding at least one alignment with a score greater than or equal to some given score. We define this quantity to be the p-value, and it can be also used to help determine if the alignment that has been returned is biologically significant. Like the e-value, as the p-value decreases, it is more likely that the alignment is biologically significant.

Slide 23 - Score, E-value and P-value compared



Score, E-value and P-value compared

Score	E-value $E(S) = K\,m\,n\,e^{-\lambda S}$	Probability $P = 1 - e^{-E(S)}$
39	12	0.999995
41	2.9	0.947456
42	1.4	0.764656
46	0.0842	0.080741
49	0.0100	0.009925
52	0.0012	0.00118
55	0.0001	0.0001

$m = 980$, $n = 10{,}030{,}834{,}086$ $(m*n \sim 10^{13})$

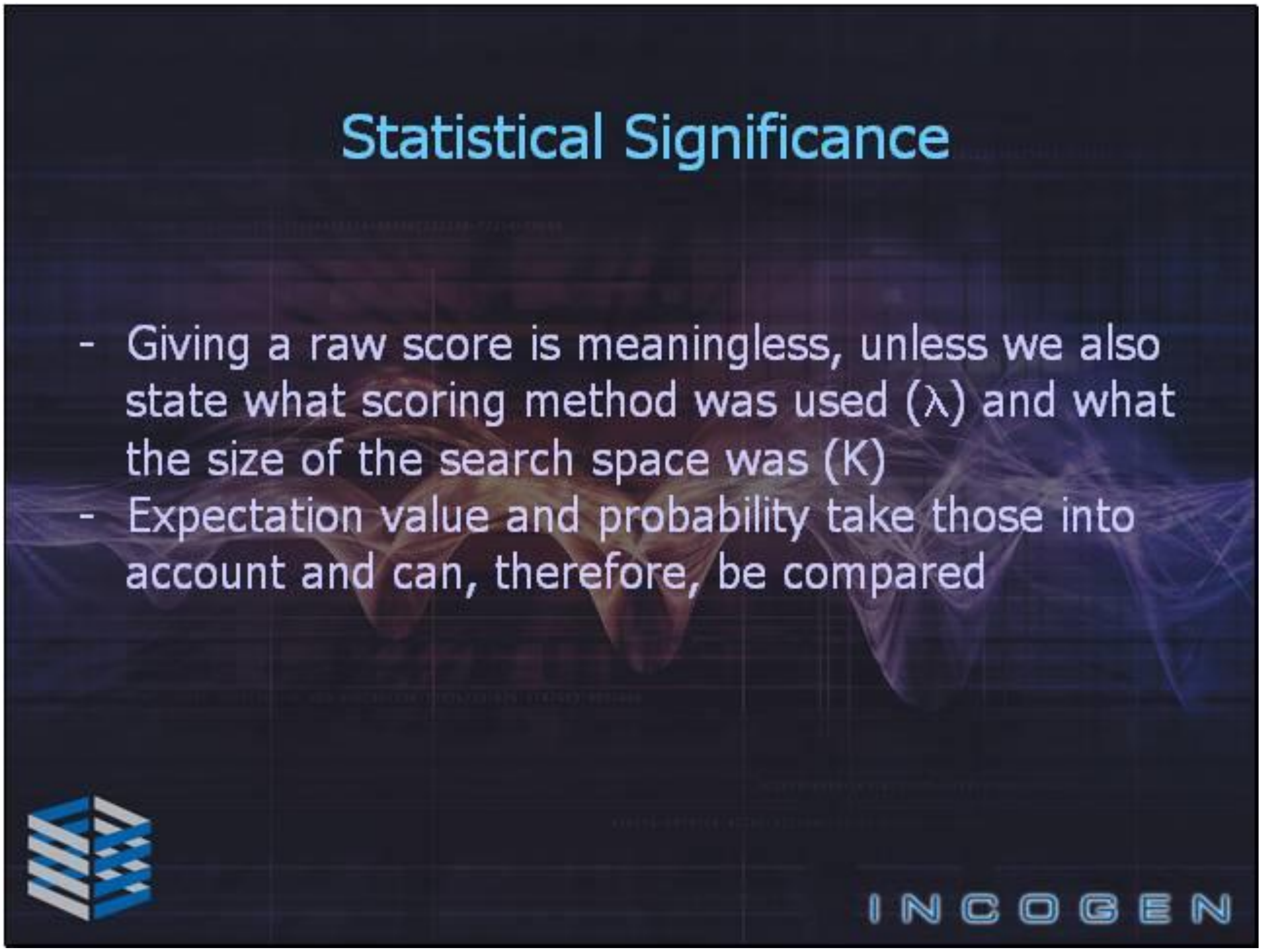
$K = 1.37$, $\lambda = 0.711$

INCOGEN

Slide notes

This table gives a brief comparison of score, e-value, and p-value. Holding all other variables constant, as the score increases, both the p-value and the e-value decrease.

Slide 24 - Statistical Significance



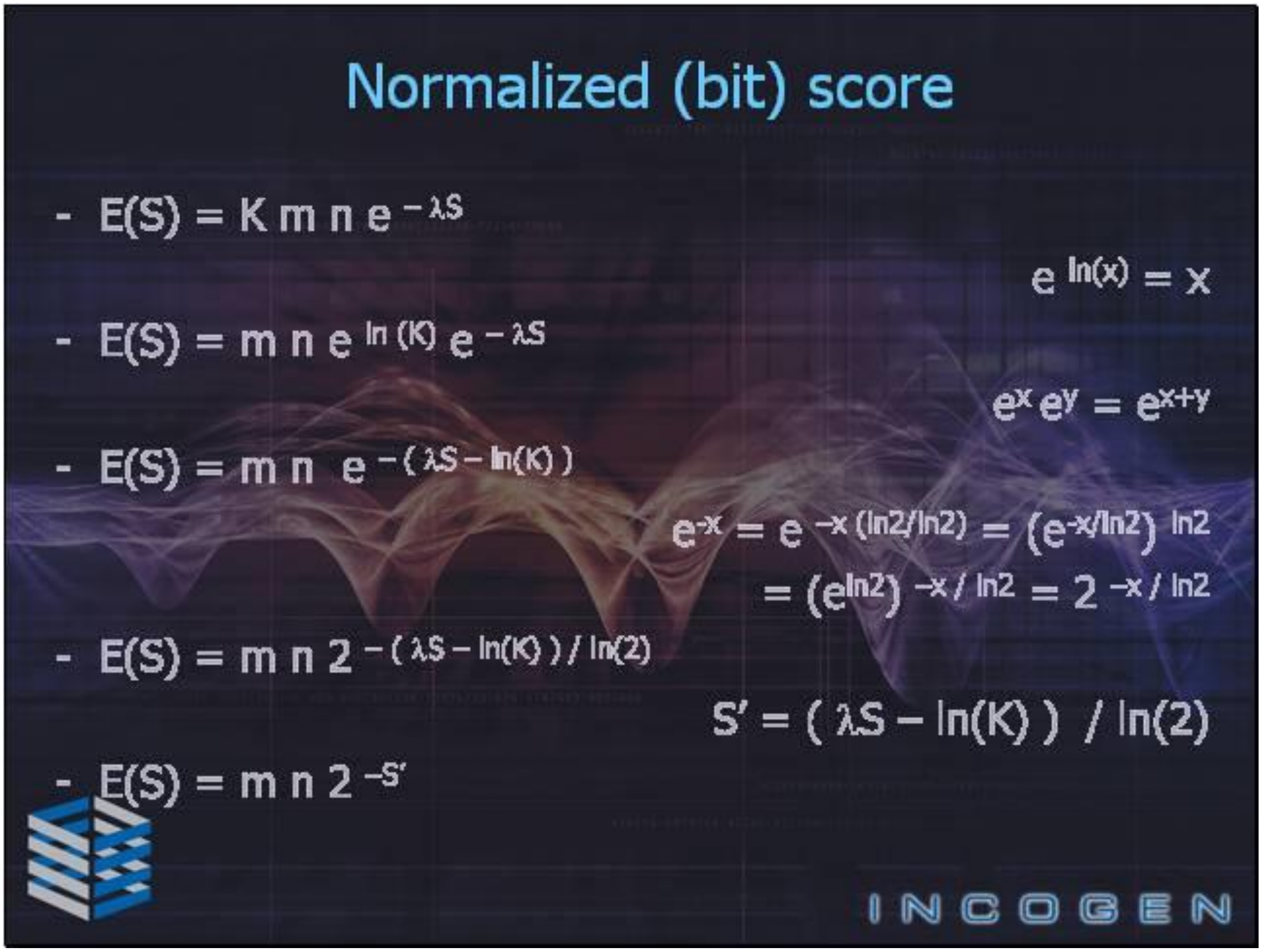
Statistical Significance

- Giving a raw score is meaningless, unless we also state what scoring method was used (λ) and what the size of the search space was (K)
- Expectation value and probability take those into account and can, therefore, be compared

Slide notes

You may be asking why not use raw score instead of e-value or p-value since the three are related and the raw score doesn’t need any additional calculations. Well, you can only directly compare the raw scores of alignments if the alignments were created using the same scoring method and against the same search space, which is the product of the database and query sequence. Therefore, the raw score is not useful in general. E-values and P-values take the search space and scoring method into account, which means that they can be directly compared to find the best alignment.

Slide 25 - Normalized (bit) score



Normalized (bit) score

- $E(S) = K\,m\,n\,e^{-\lambda S}$
- $E(S) = m\,n\,e^{\ln(K)}\,e^{-\lambda S}$
- $E(S) = m\,n\;e^{-(\lambda S - \ln(K))}$
- $E(S) = m\,n\,2^{-(\lambda S - \ln(K))/\ln(2)}$
- $E(S) = m\,n\,2^{-S'}$

$$e^{\ln(x)} = x$$

$$e^{x} e^{y} = e^{x+y}$$

$$e^{-x} = e^{-x(\ln 2/\ln 2)} = (e^{-x/\ln 2})^{\ln 2}$$
$$= (e^{\ln 2})^{-x/\ln 2} = 2^{-x/\ln 2}$$

$$S' = (\lambda S - \ln(K))\;/\;\ln(2)$$

INCOGEN

Slide notes

Another useful value used to measure the strength of a pairwise alignment is the bit score, S', derived in this slide.

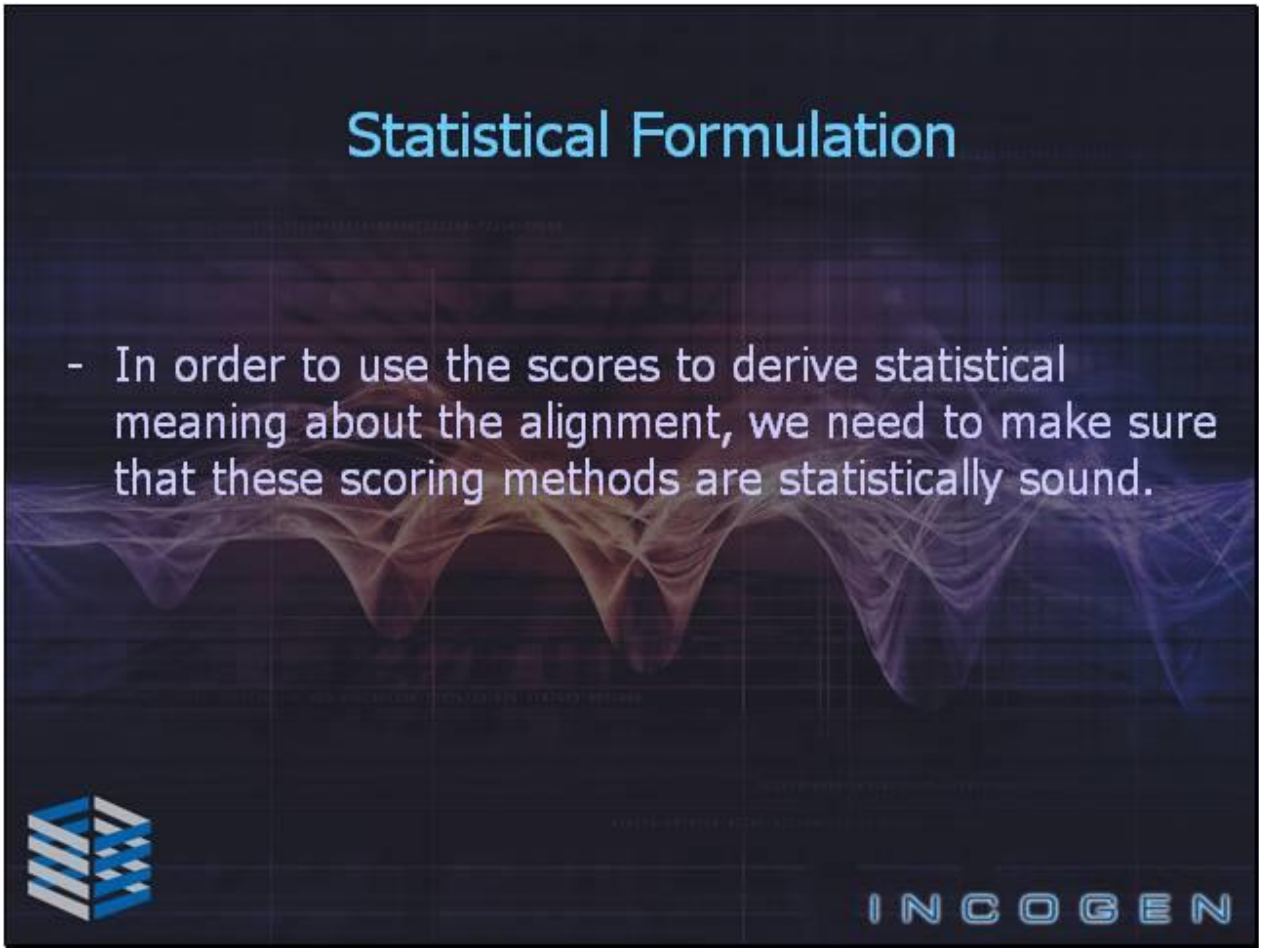
Slide 26 - Statistical Formulation

Statistical Formulation

Slide notes

And now, a few words about the statistics behind these models.

Slide 27 - Statistical Formulation



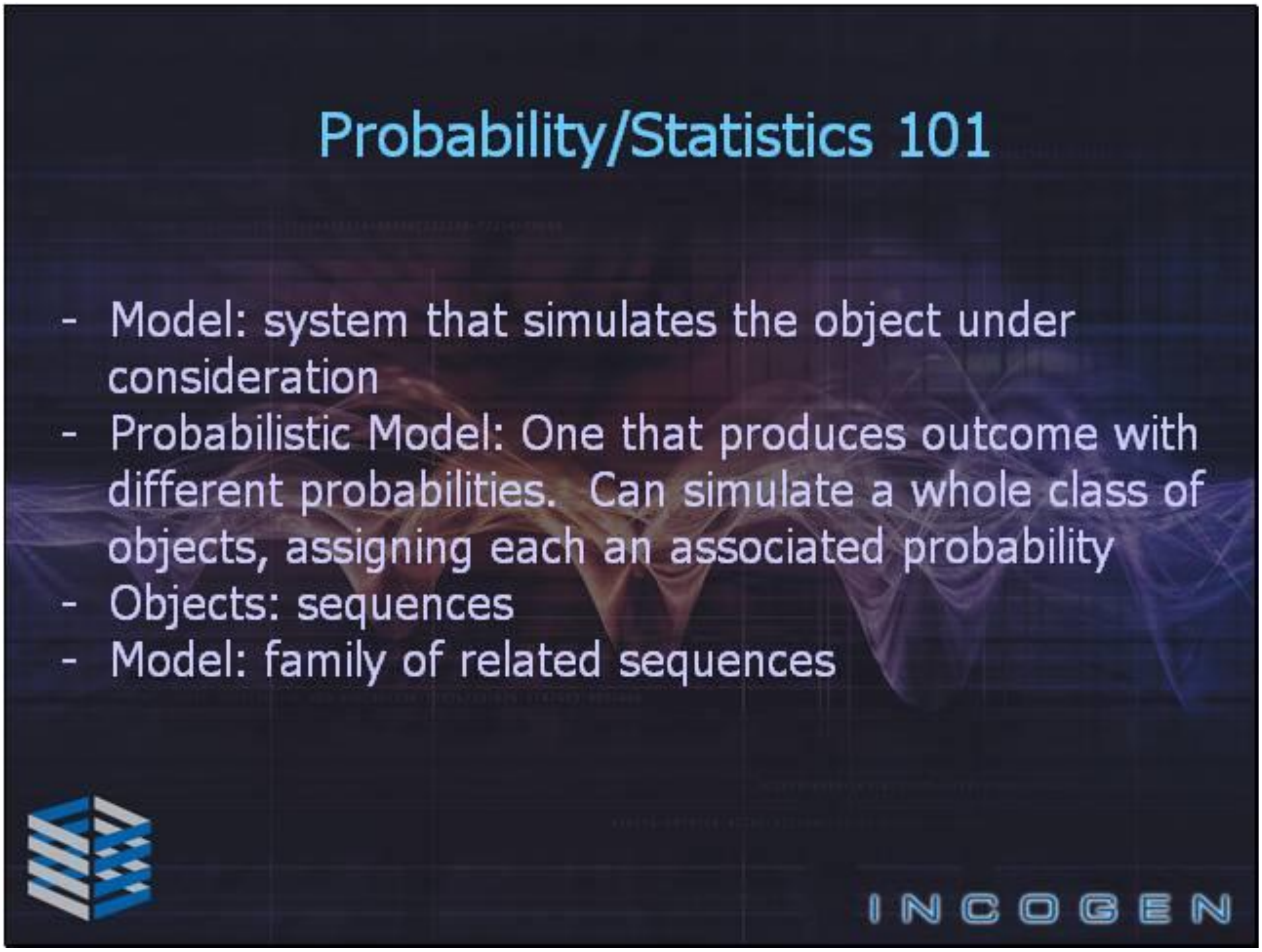
Statistical Formulation

- In order to use the scores to derive statistical meaning about the alignment, we need to make sure that these scoring methods are statistically sound.

Slide notes

In order for the scores to be useful, we need to make sure that the scoring methods we use are statistically valid.

Slide 28 - Probability/Statistics 101



Probability/Statistics 101

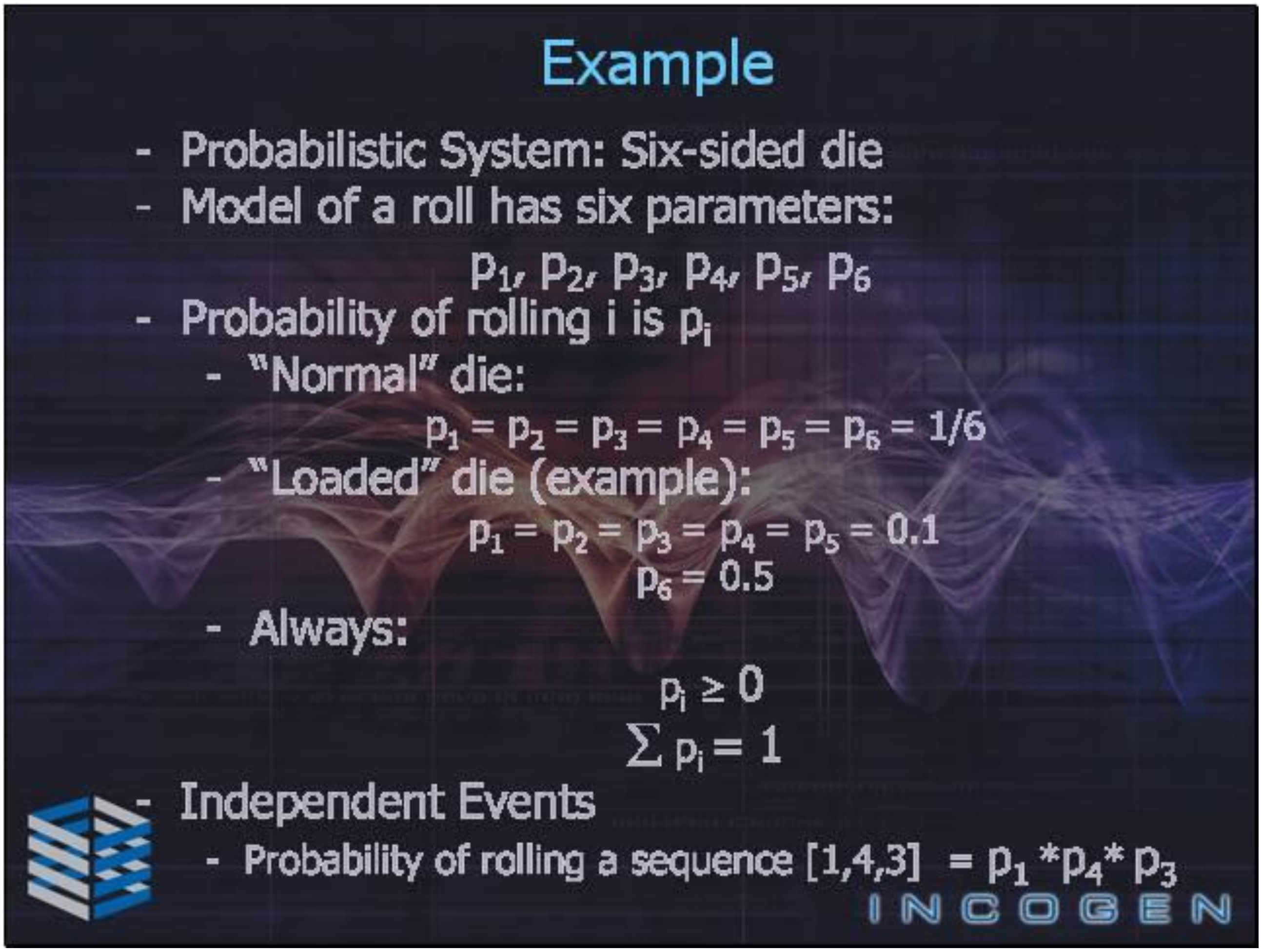
- Model: system that simulates the object under consideration
- Probabilistic Model: One that produces outcome with different probabilities. Can simulate a whole class of objects, assigning each an associated probability
- Objects: sequences
- Model: family of related sequences

INCOGEN

Slide notes

Let’s go over a few terms used frequently in statistics. A model is a system that simulates the object under consideration. A probabilistic model is a model that produces different outcomes based on a set of probabilities. In our case, the objects we are simulating are sequences, and the model is a family of related sequences.

Slide 29 - Example



Example

- Probabilistic System: Six-sided die
- Model of a roll has six parameters:

$$p_1, p_2, p_3, p_4, p_5, p_6$$

- Probability of rolling i is p_i
  - “Normal” die:

$$p_1 = p_2 = p_3 = p_4 = p_5 = p_6 = 1/6$$

  - “Loaded” die (example):

$$p_1 = p_2 = p_3 = p_4 = p_5 = 0.1$$
$$p_6 = 0.5$$

  - Always:

$$p_i \geq 0$$
$$\sum p_i = 1$$

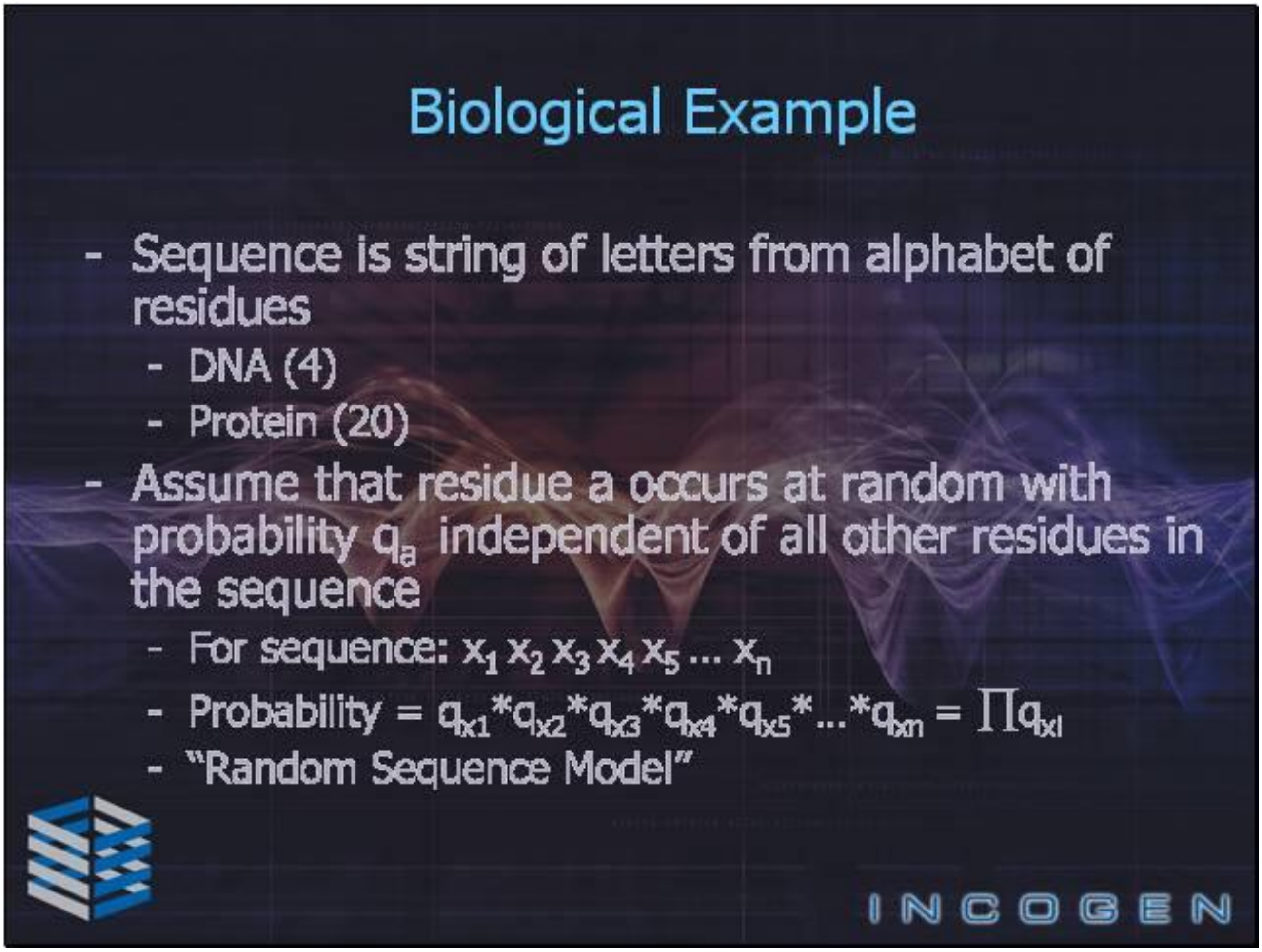
- Independent Events
  - Probability of rolling a sequence [1,4,3] = $p_1 * p_4 * p_3$

INCOGEN

Slide notes

Let’s look at a more general example-that of a six-sided die. There are 6 outcomes for each roll of a die, each with their own probability. For a “normal” die, the probability of any given number appearing after a roll is the same: 1/6. If the die is “loaded,” the probabilities of the six outcomes is not the same. Regardless of whether the die is “normal” or “loaded,” the probability of any outcome must be greater than or equal to zero and the sums of the probabilities of all of the outcomes must equal 1. If the events we are modeling are independent of each other, such as rolls of a die, then we can calculate the probability of a sequence of events by multiplying together the probabilities of the events occurring separately.

Slide 30 - Biological Example



Biological Example

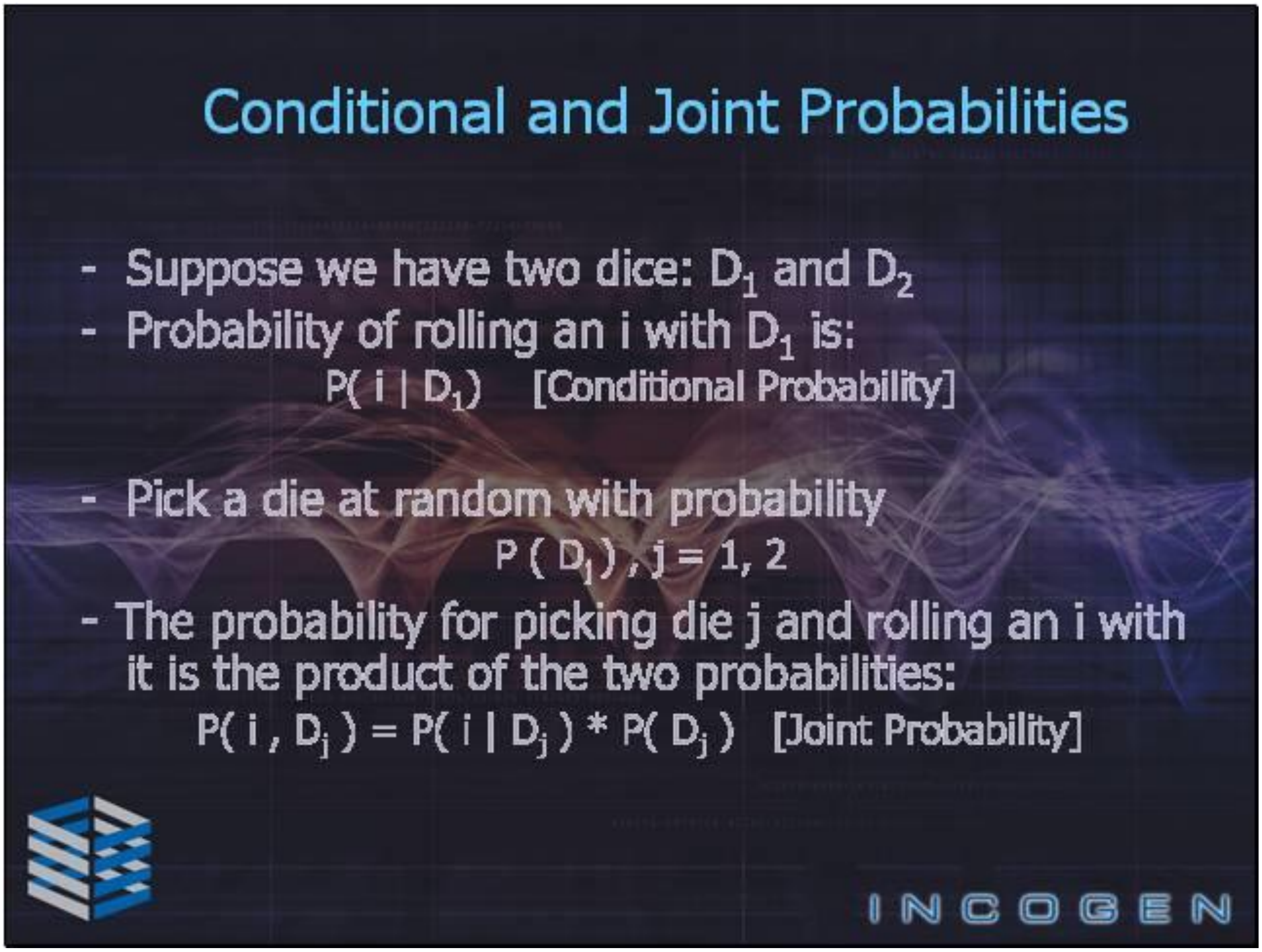
- Sequence is string of letters from alphabet of residues
  - DNA (4)
  - Protein (20)
- Assume that residue a occurs at random with probability q_a independent of all other residues in the sequence
  - For sequence: $x_1 x_2 x_3 x_4 x_5 \ldots x_n$
  - Probability = $q_{x1} * q_{x2} * q_{x3} * q_{x4} * q_{x5} * \ldots * q_{xn} = \prod q_{xl}$
  - "Random Sequence Model"

INCOGEN

Slide notes

Now, let's look at a biological example of an independent system: a DNA or protein sequence. A sequence is built from an "alphabet" of residues. A DNA sequence's alphabet contains 4 letters, while a protein sequence's alphabet contains 20. To create a random sequence, we can assume that each residue occurs with a certain probability, regardless of the other residues in the sequence. This allows us to calculate the probability of a particular sequence by multiplying together the probabilities of all of the residues in the sequence. This is called the "Random Sequence Model."

Slide 31 - Conditional and Joint Probabilities



Conditional and Joint Probabilities

- Suppose we have two dice: D_1 and D_2
- Probability of rolling an i with D_1 is:

  $P(i \mid D_1)$ [Conditional Probability]

- Pick a die at random with probability

  $P (D_1) , j = 1, 2$

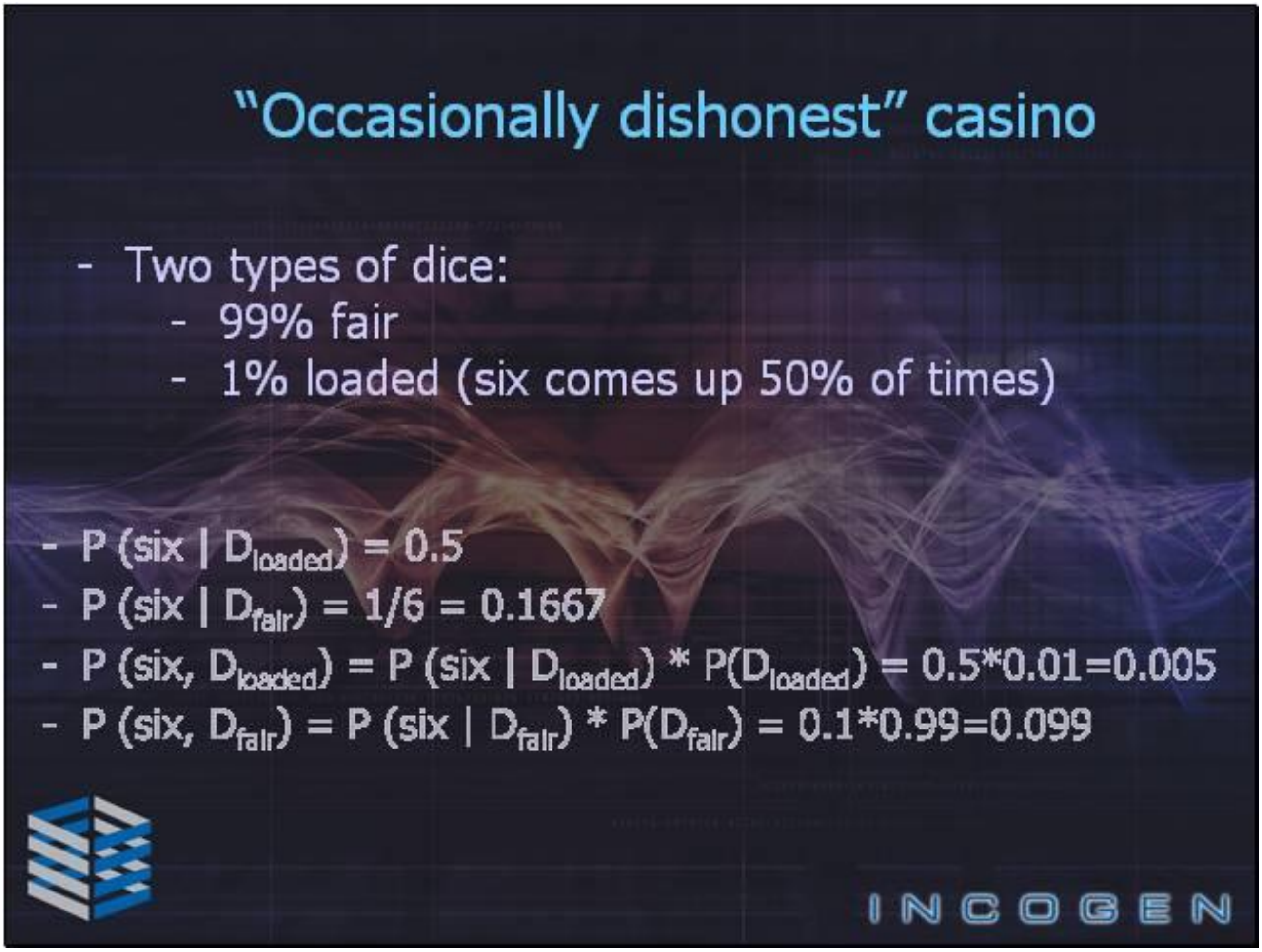
- The probability for picking die j and rolling an i with it is the product of the two probabilities:

  $P(i , D_j) = P(i \mid D_j) * P(D_j)$ [Joint Probability]

Slide notes

Let’s go back to our previous example involving a die, except now, we are going to look at a system that contains two dice. Each of these dice could have a different probability of rolling a given number, so we must specify both the die and the outcome of the roll in order to find the probability; this is called the conditional probability. (For example, what is the probability of rolling a 2 given the condition that you are rolling the red die?) If the condition isn’t guaranteed, then the probability depends on meeting that condition and then getting the outcome; this is called a joint probability. (For example, what is the probability of rolling a 2 on a red die if you pick from a bag that contains both red and blue dice?) To calculate a joint probability, you multiply the probability of meeting the condition with the probability of getting the outcome once the condition is met.

Slide 32 - “Occasionally dishonest” casino



“Occasionally dishonest” casino

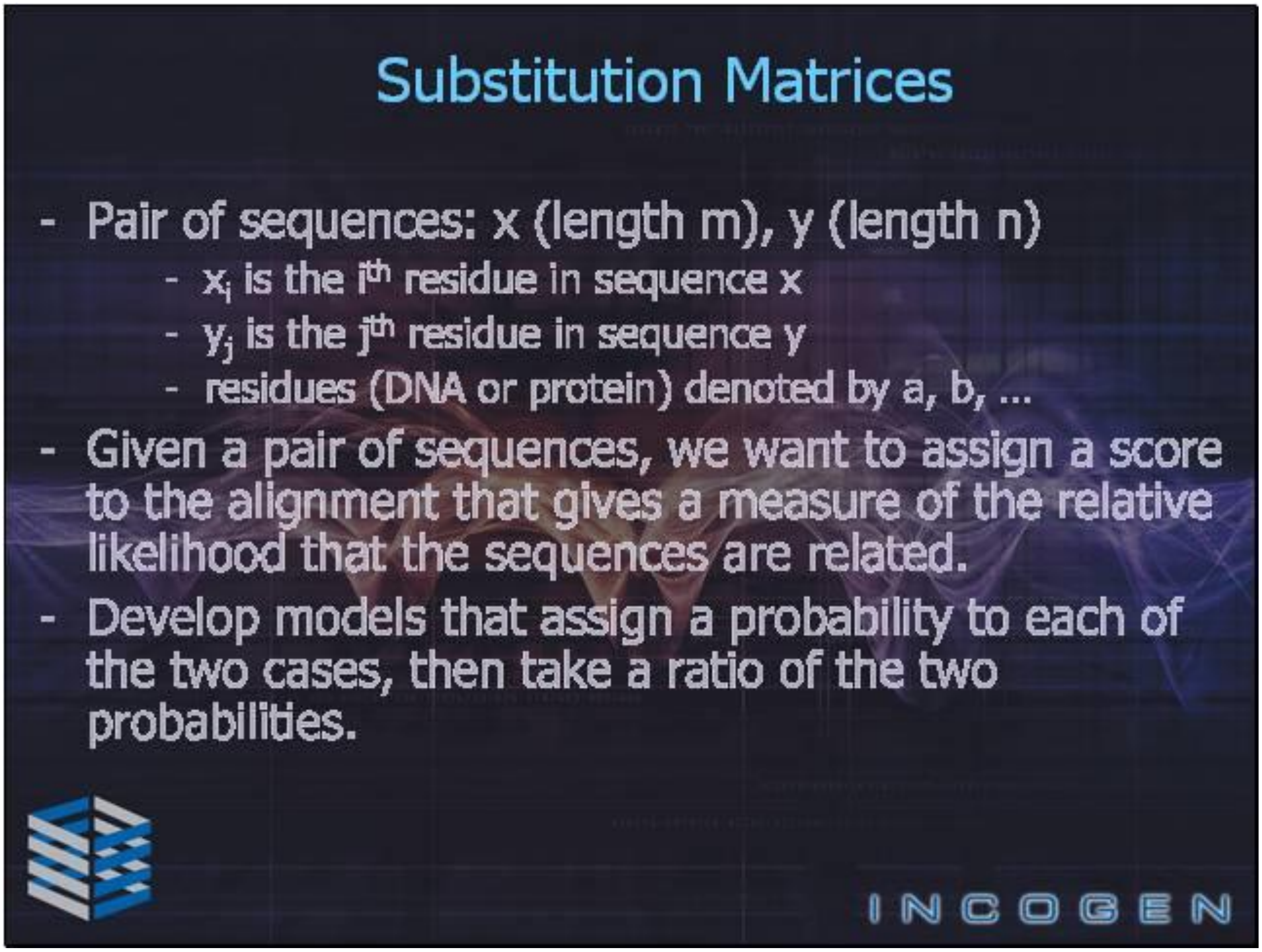
- Two types of dice:
  - 99% fair
  - 1% loaded (six comes up 50% of times)

- $P\ (\text{six} \mid D_{\text{loaded}}) = 0.5$
- $P\ (\text{six} \mid D_{\text{fair}}) = 1/6 = 0.1667$
- $P\ (\text{six}, D_{\text{loaded}}) = P\ (\text{six} \mid D_{\text{loaded}}) * P(D_{\text{loaded}}) = 0.5*0.01=0.005$
- $P\ (\text{six}, D_{\text{fair}}) = P\ (\text{six} \mid D_{\text{fair}}) * P(D_{\text{fair}}) = 0.1*0.99=0.099$

Slide notes

Let’s look at another concrete example. Let’s say, for instance,. That you are in an “occasionally dishonest” casino where 99% of the dice are fair and 1% are “loaded” so that the probability of rolling a 6 is 50%. So, we know that the probability of rolling a 6 on a “loaded” die is 50%, and the probability of rolling a 6 on a “normal” die is approximately 17%. To find the general probability of rolling a 6, we must take into account both the normal dice and the loaded dice. To do this, we multiply the probability of getting a “loaded” die by the probability of rolling a 6 with a “loaded” die (50%). We do the same for the "normal” die, and then sum the two probabilities together.

Slide 33 - Substitution Matrices



Substitution Matrices

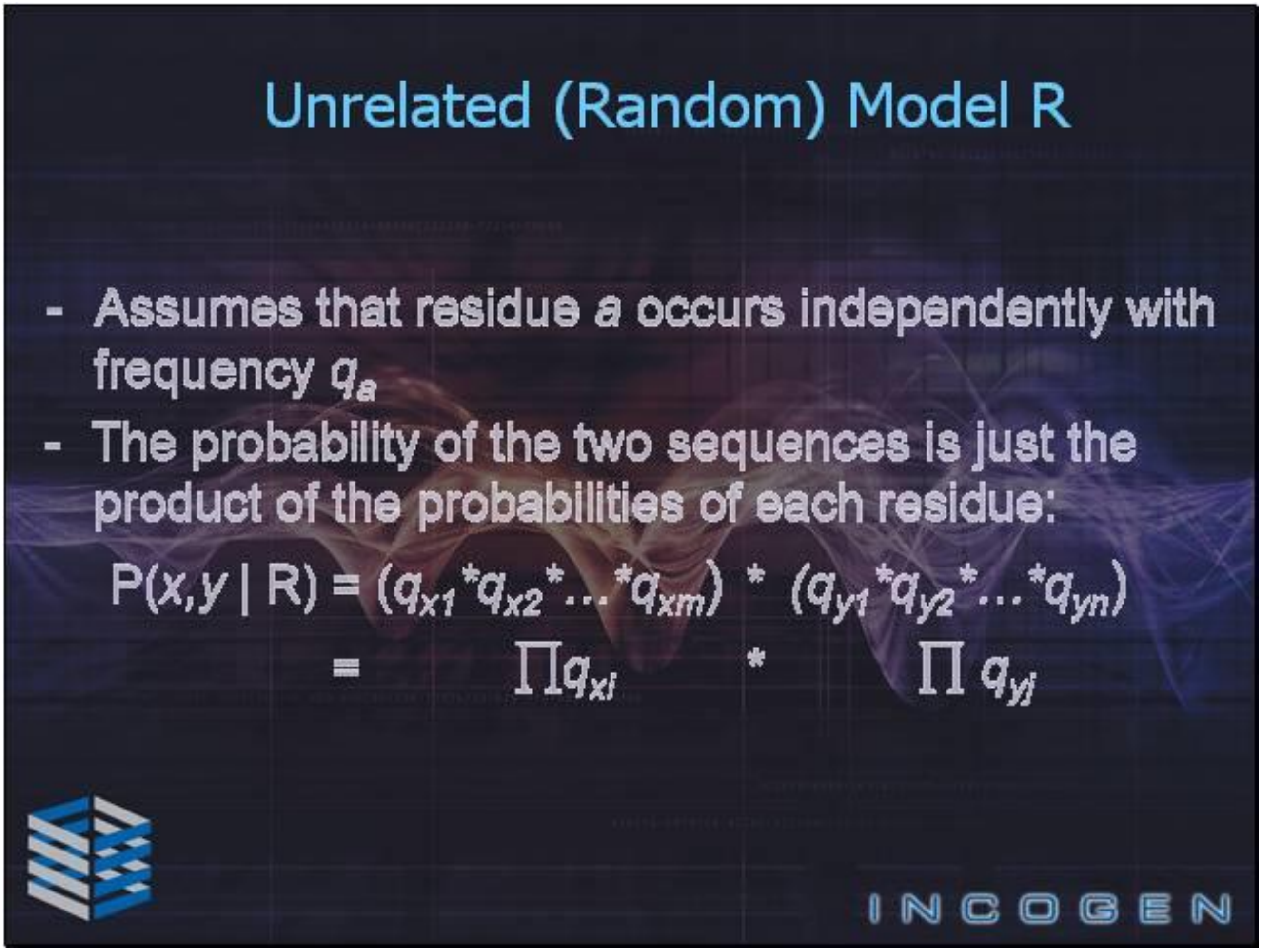
- Pair of sequences: x (length m), y (length n)
  - x_i is the i^{th} residue in sequence x
  - y_j is the j^{th} residue in sequence y
  - residues (DNA or protein) denoted by a, b, ...
- Given a pair of sequences, we want to assign a score to the alignment that gives a measure of the relative likelihood that the sequences are related.
- Develop models that assign a probability to each of the two cases, then take a ratio of the two probabilities.

INCOGEN

Slide notes

Let’s look a little at how this relates to substitution matrices. The general issue: given a pair of sequences, we want to assign a score to an alignment that gives us some idea of the relative likelihood that the two sequences are related versus unrelated. To do that, we need to develop models for the related and unrelated cases that calculate the probability of each case and then take a ratio of the two probabilities.

Slide 34 - Unrelated (Random) Model R



Unrelated (Random) Model R

- Assumes that residue *a* occurs independently with frequency q_a
- The probability of the two sequences is just the product of the probabilities of each residue:

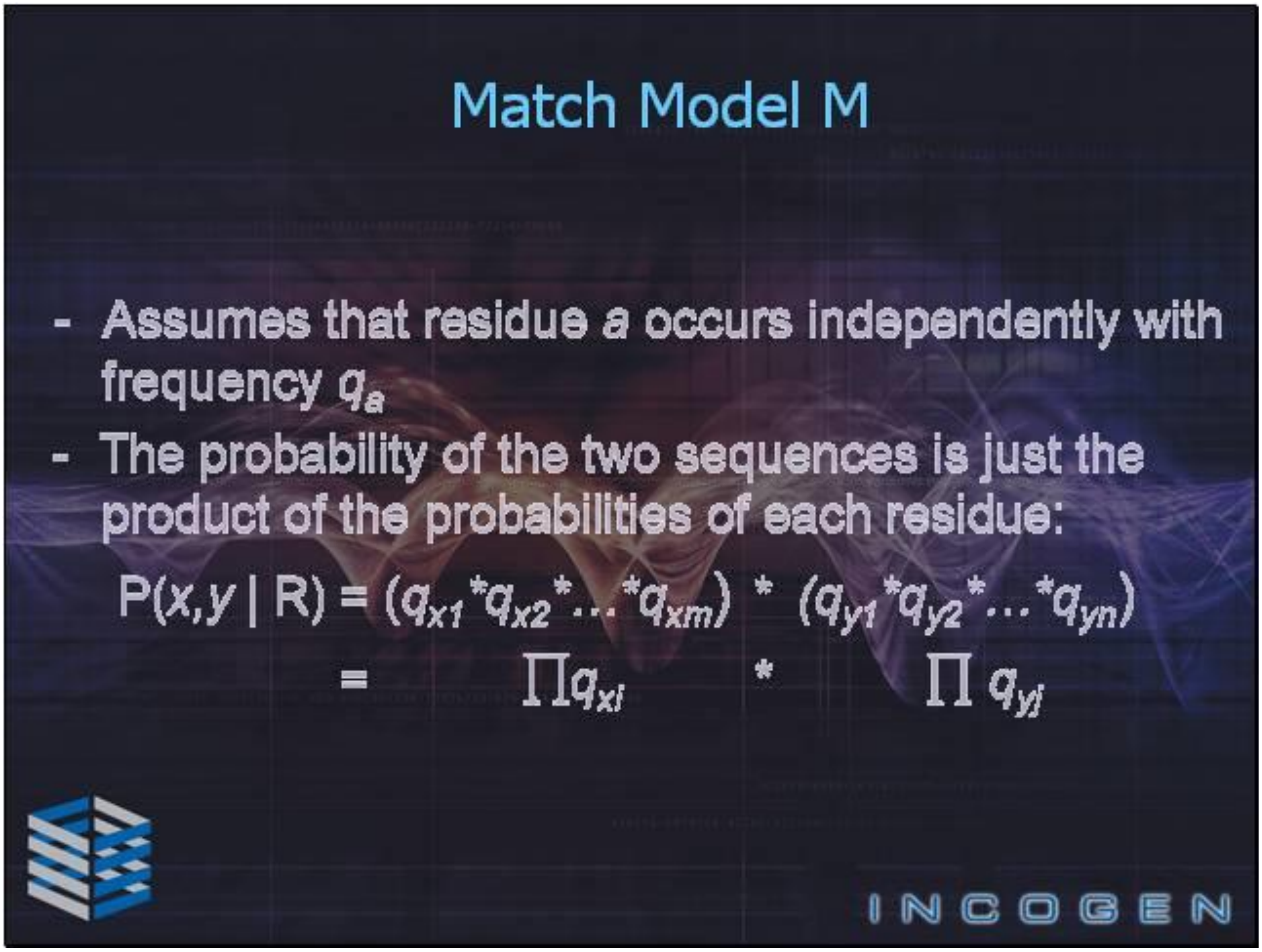
$$P(x,y \mid R) = (q_{x1} * q_{x2} * \ldots * q_{xm}) \; * \; (q_{y1} * q_{y2} * \ldots * q_{yn})$$

$$= \prod q_{xi} \quad * \quad \prod q_{yj}$$

Slide notes

First, let’s look at the model for unrelated sequences. In this model, we assume that each residue occurs independently of any other in the sequence with some given probability. The probability that the two sequences exist independently of each other is just the product of probabilities of each residue in each sequence.

Slide 35 - Match Model M



Match Model M

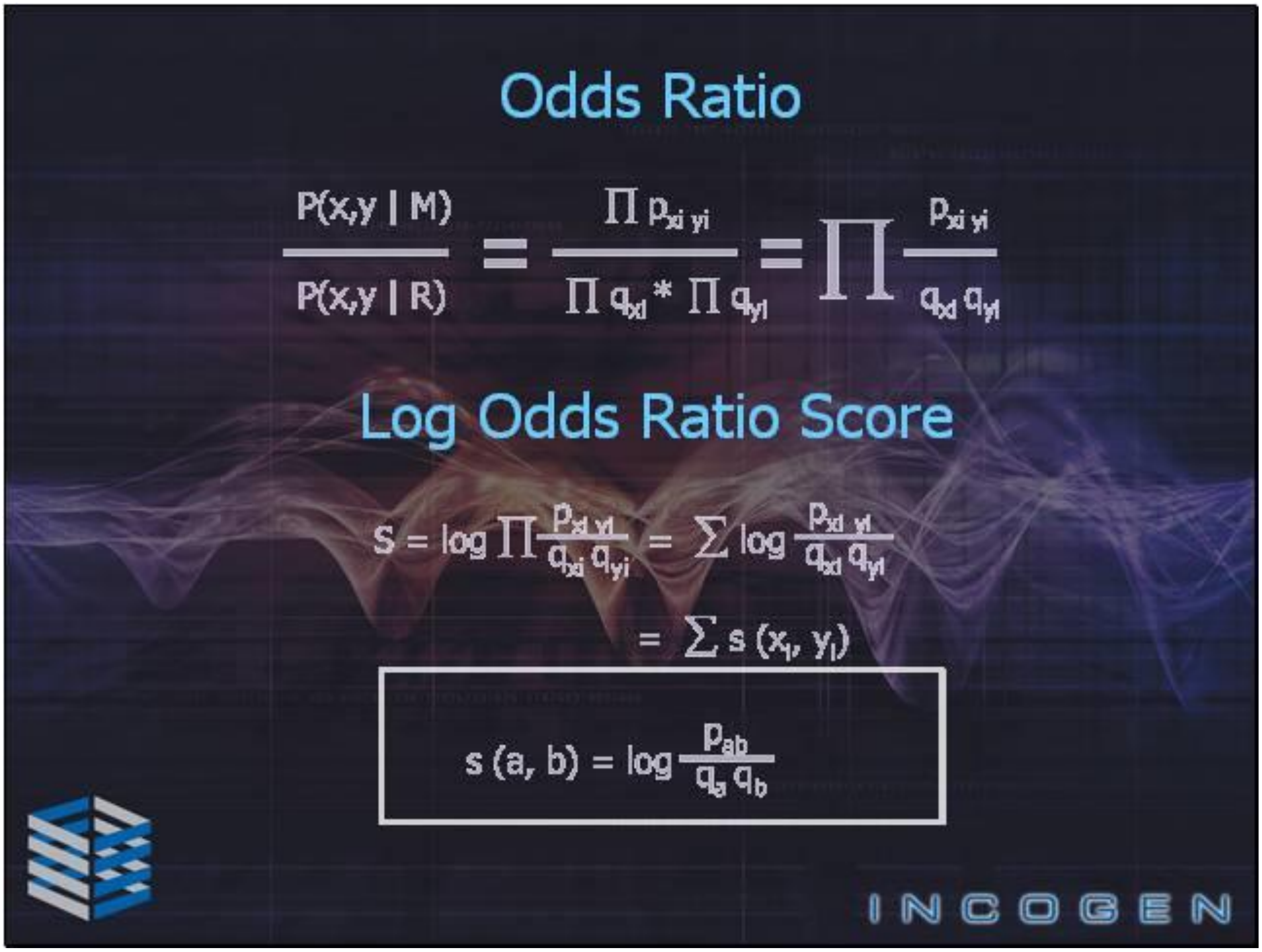
- Assumes that residue *a* occurs independently with frequency q_a
- The probability of the two sequences is just the product of the probabilities of each residue:

$$P(x,y \mid R) = (q_{x1} * q_{x2} * \ldots * q_{xm}) \; * \; (q_{y1} * q_{y2} * \ldots * q_{yn})$$
$$= \prod q_{xi} \; * \; \prod q_{yj}$$

Slide notes

Now let’s look at the model for related sequences. In this case, each aligned pair of residues occurs with some joint probability. This joint probability represents the idea that the two residues are derived from some unknown common ancestor. The probability that these aligned pairs exist is given by the product of all of the joint probabilities.

Slide 36 - Odds Ratio



Odds Ratio

$$\frac{P(x,y \mid M)}{P(x,y \mid R)} = \frac{\prod p_{x_i y_i}}{\prod q_{x_i} * \prod q_{y_i}} = \prod \frac{p_{x_i y_i}}{q_{x_i} q_{y_i}}$$

Log Odds Ratio Score

$$S = \log \prod \frac{p_{x_i y_i}}{q_{x_i} q_{y_i}} = \sum \log \frac{p_{x_i y_i}}{q_{x_i} q_{y_i}}$$

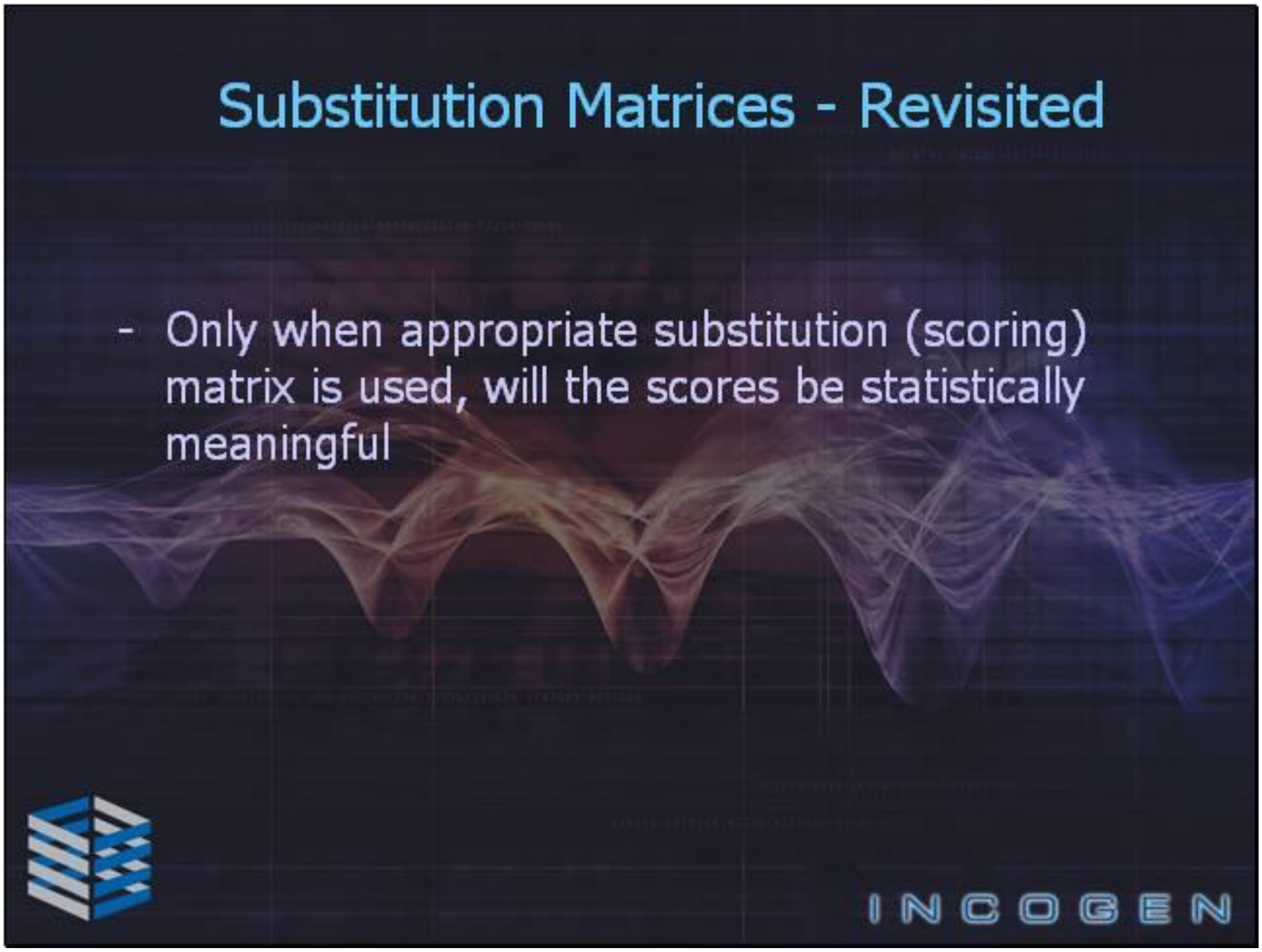
$$= \sum s(x_i, y_i)$$

$$\boxed{s(a, b) = \log \frac{p_{ab}}{q_a q_b}}$$

Slide notes

To get these probabilities into a form that can be combined by adding, we need to find the log-odds ratio score. First, we find the odds ratio by dividing the probability that the two sequences are related by the probability that the two sequences exist independently of each other. We then take the log of that to find the log-odd ratio score. Since each aligned pair is independent of the other aligned pairs in the sequences, we can also calculate a log-odds ratio for the aligned pairs independently. It is this value that appears in scoring matrices.

Slide 37 - Substitution Matrices - Revisited



Substitution Matrices - Revisited

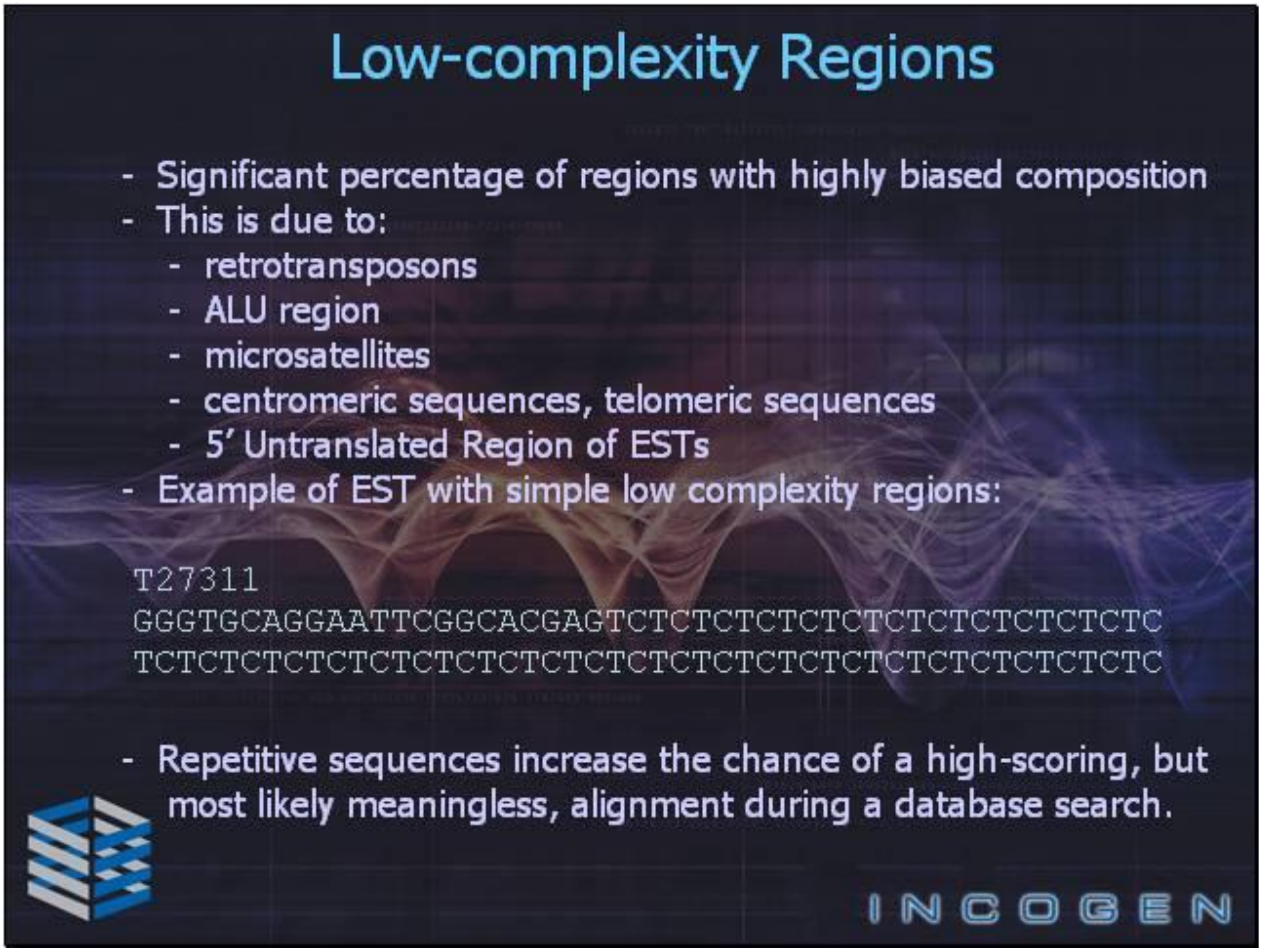
- Only when appropriate substitution (scoring) matrix is used, will the scores be statistically meaningful

INCOGEN

Slide notes

To reiterate, scores are only statistically meaningful when the appropriate scoring matrix is used.

Slide 38 - Low-complexity Regions



Low-complexity Regions

- Significant percentage of regions with highly biased composition
- This is due to:
  - retrotransposons
  - ALU region
  - microsatellites
  - centromeric sequences, telomeric sequences
  - 5' Untranslated Region of ESTs
- Example of EST with simple low complexity regions:

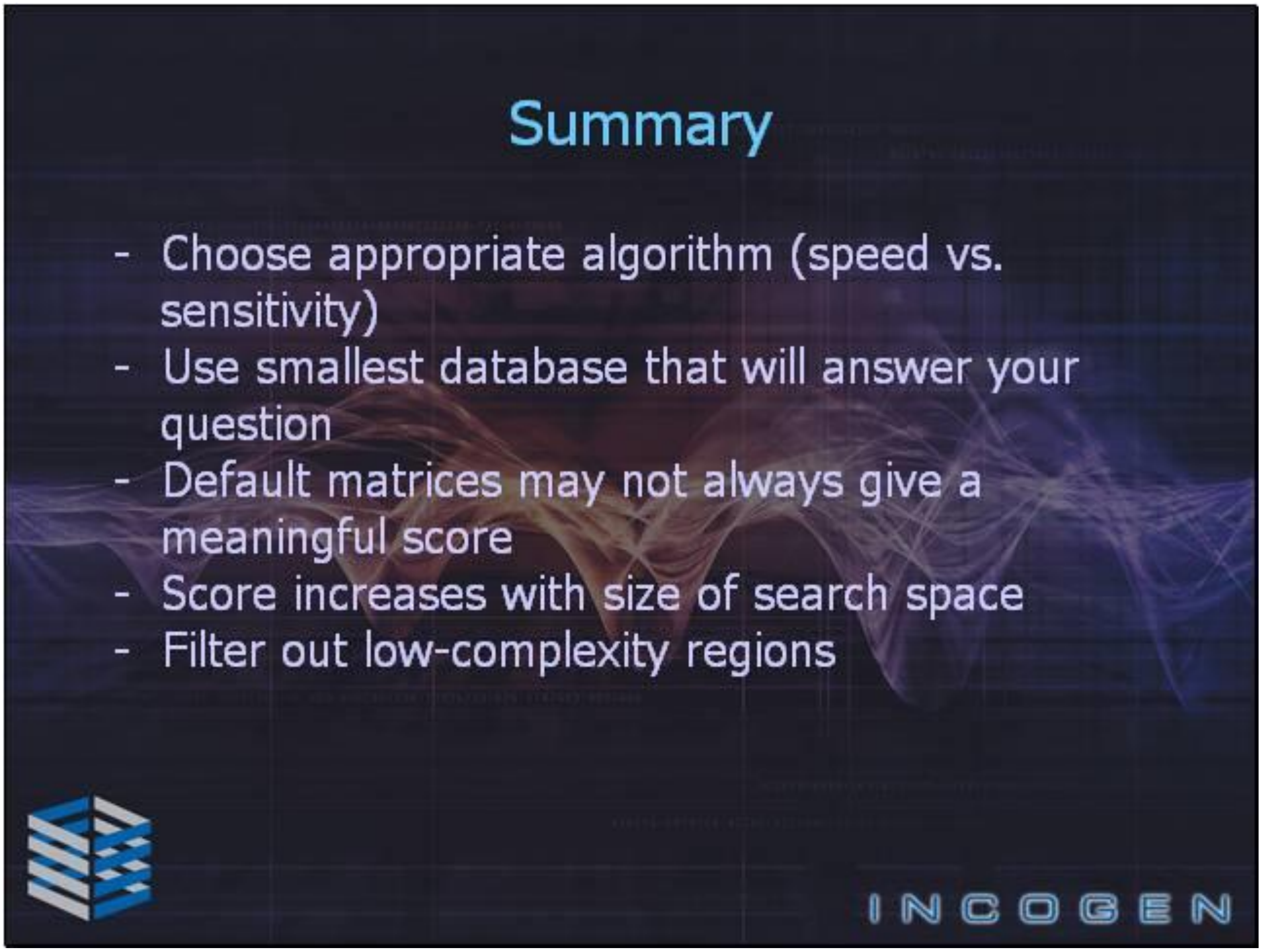
```
T27311
GGGTGCAGGAATTCGGCACGAGTCTCTCTCTCTCTCTCTCTCTCTCTC
TCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTC
```

- Repetitive sequences increase the chance of a high-scoring, but most likely meaningless, alignment during a database search.

Slide notes

The last thing we'll discuss in this lecture is the effect of low-complexity regions on alignments and scoring. Some sequences contain regions that are highly repetitive and low-complexity. This can occur because of a variety of reasons. These repetitive regions increase the chance of a high-scoring, but meaningless, alignment during database searches. Therefore, it is better to mask these regions before beginning a pairwise search.

Slide 39 - Summary



Summary

- Choose appropriate algorithm (speed vs. sensitivity)
- Use smallest database that will answer your question
- Default matrices may not always give a meaningful score
- Score increases with size of search space
- Filter out low-complexity regions

INCOGEN

Slide notes

In summary, you must take speed and sensitivity into account when choosing the appropriate algorithm for determining alignments. To further speed up the algorithms, use the smallest database that will answer your question. Take some thought when picking which scoring matrix to use; the default matrix may not provide you with meaningful results. The score of any alignment increases with the size of the search, which is why it is advisable to use p-values and e-values for comparison purposes. Finally, filter out or mask low-complexity regions in sequences to prevent spurious, but high scoring, alignments from appearing in your results.