

**Scoring matrices
and the
statistical significance
of
molecular sequence features**

Antoine van Kampen
Bioinformatics Laboratory
a.h.vankampen@amc.uva.nl
<http://bioinformatica.amc.uva.nl>

Content

1. Application of scoring matrices
2. Sequence alignment & pattern finding
3. PAM (and BLOSUM) scoring matrices
4. Statistical Evaluation of scores
5. Scores and information theory
6. BLAST

Options for advanced blasting

[Limit by entrez query](#) or select from:

[Composition-based statistics](#)

[Choose filter](#) Low complexity Mask for lookup table only Mask lower case

[Expect](#)

Expectation value

[Word Size](#)

[Matrix](#) Gap Costs

Scoring matrix

[PSSM](#)

[Other advanced](#)

[PHI pattern](#)

Format

You will learn to understand BLAST output

```
output>gi|6006425|emb|CAB56829.1| hemoglobin alpha chain
```

```
Length = 142
```

```
Score = 33.9 bits (76), Expect = 0.66
```

```
Identities = 15/15 (100%), Positives = 15/15 (100%)
```

```
Query: 1  MVLSAADKGNVKA AW 15
        MVLSAADKGNVKA AW
Sbjct: 1  MVLSAADKGNVKA AW 15
```

Database: All non-redundant GenBank CDS

Number of letters in database: 436,700,696

Number of sequences in database: 1,364,053

Gapped

Lambda K H

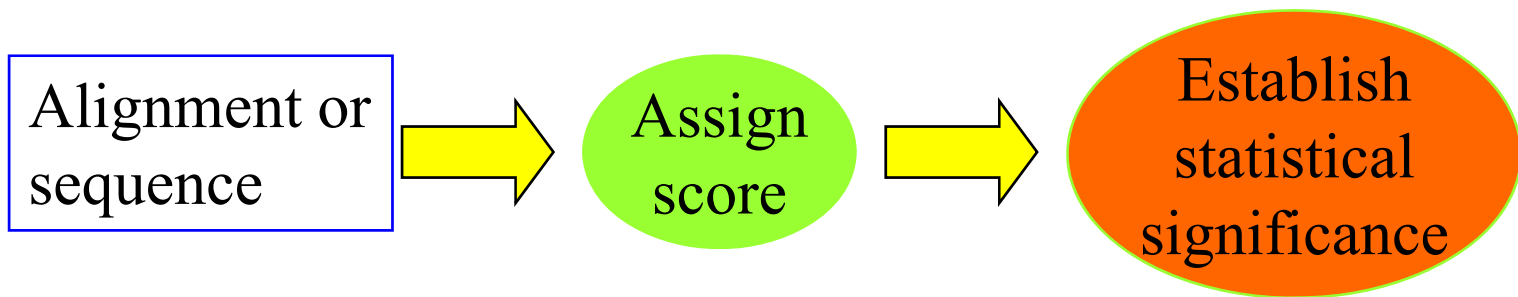
0.267

0.0410

0.140

Sequence analysis

- **Sequence alignment** of two or more sequences
 - gives information about: function, evolutionary history
- Finding **patterns** in sequences
 - e.g., transmembrane regions, potential glycosylation sites



Example: pattern finding

Consider protein sequence:

XXXXXXXXXXXX **TTTTTTTTTTTTTT**XXXXXXXXXXXXXXXXXXXX Score=100

XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX **CCCCCCCCCC**xxx Score=90

T=transmembrane region

C=chance segment (e.g. random distribution of amino acids)

- 1) How can we define a scoring scheme that distinguishes between T and C regions ?
- 2) How can we establish whether a score (100 and 90) is statistical significant?

Example: sequence alignment

A HBA_HUMAN GSAQVKGHGKKVADALTNAVAHVDI
 G+ +VK+HGKKV A++++AH+D+
 HBB_HUMAN GNPKVKAHGKKVLGAFSDGLAHLDM

B HBA_HUMAN GSAQVKGHGKKVADALTNAVAHV--
 ++ ++++H+ KV + +A ++
 LGB2_LUPLU NNPELQAHAGKVFKLVEAAIQIQVIGVY+D+LHNEGQVHYDRO
(leghaemoglobin from yellow lupin: same 3D-structure; b

C HBA_HUMAN GSAQVKGHGKKVADALTNAVAHVDDMPN
 GS+ + G + +D L ++ H+ D+
 F11G11.2 GS ANAALLDEF
(nematode g e)



Example: sequence alignment

A

```
HBA_HUMAN      GSAQVKGHGKKVADALTNAVAHVDDMPNALSALSDLHAAHKL
                G+ +VK+HGKKV  A++++AH+D++ +++++LS+LH  KL
HBB_HUMAN      GNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKL
```

B

```
HBA_HUMAN      GSAQVKGHGKKVADALTNAVAHV---D--DMPNALSALSDLHAAHKL
                ++ +++++H+ KV  + +A ++                +L+ L+++H+ K
LGB2_LUPLU     NNPELQAHAGKVFKLVYEAAIQLQVTGVVVTDATLKNLGSVHVSKG
(leghaemoglobin from yellow lupin: same 3D-structure; both oxygen binding)
```

C

```
HBA_HUMAN      GSAQVKGHGKKVADALTNAVAHVDDMPNALSALSSD----LHAAHKL
                GS+ + G +   +D L  ++ H+ D+  A +AL D   ++AH+
F11G11.2      GSGYLVGDSLTFVDLL--VAQHTADLLAANAALLDEFFPQFKAHQE
(nematode glutathione S-transferase homologue )
```

Alignment (b) and (c) have similar amount of 'positives'. However, alignment (b) represents true biological relationship while (c) is spurious high-scoring alignment.

Scoring matrices

PAM and BLOSUM

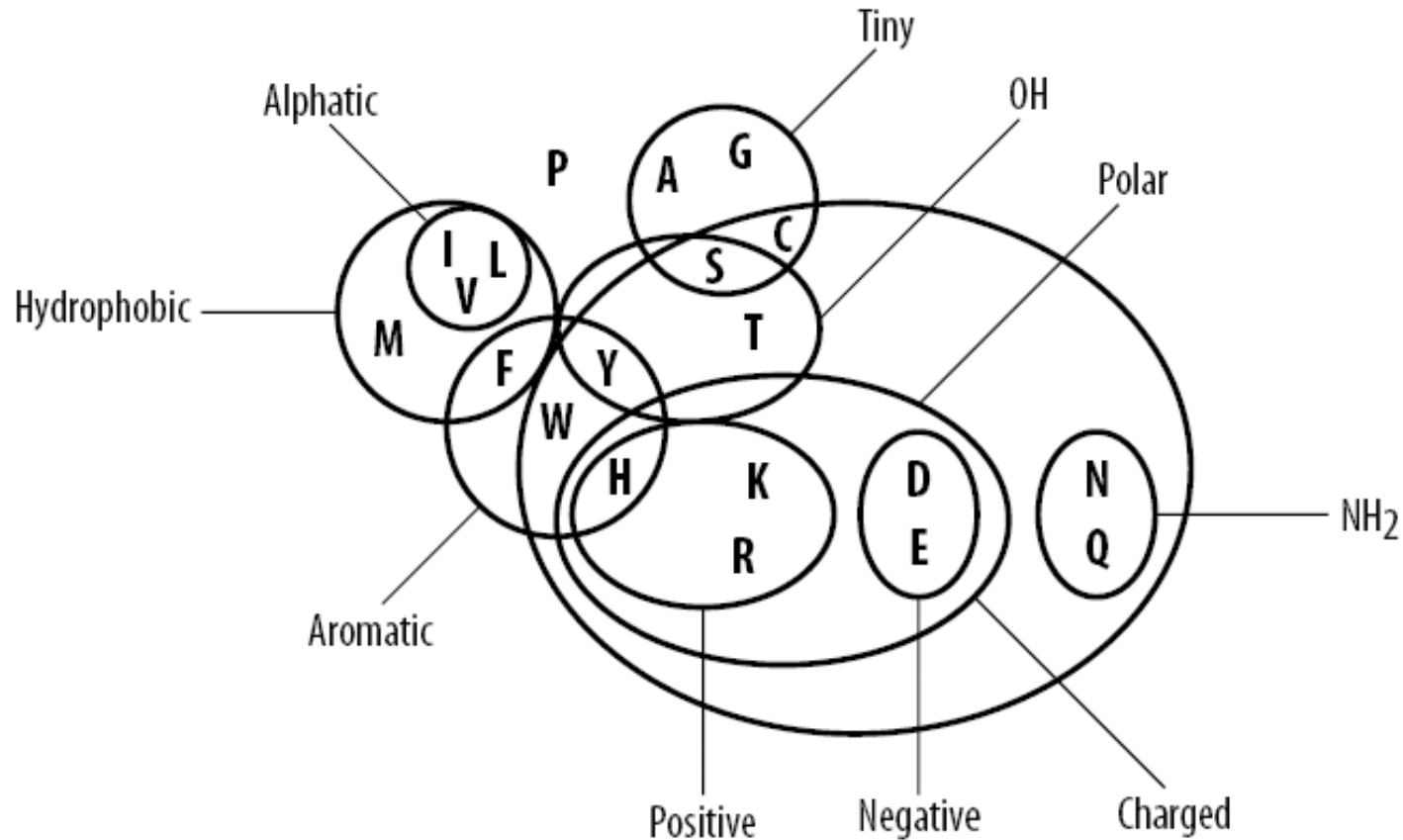
Scoring matrices 1

- Comparing sequences: simple scoring schemes such as
match=+1, mismatch=0, space=-1
are not sufficient.

| | | | | | | | | | | | | |
|-----------|---|---|---|---|---|---|---|---|---|---|---|---|
| HBA_HUMAN | G | S | A | Q | V | K | G | H | G | K | K | V |
| HBB_HUMAN | G | N | P | K | V | K | A | H | G | K | K | V |
| Score | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 |

Scoring matrices 2

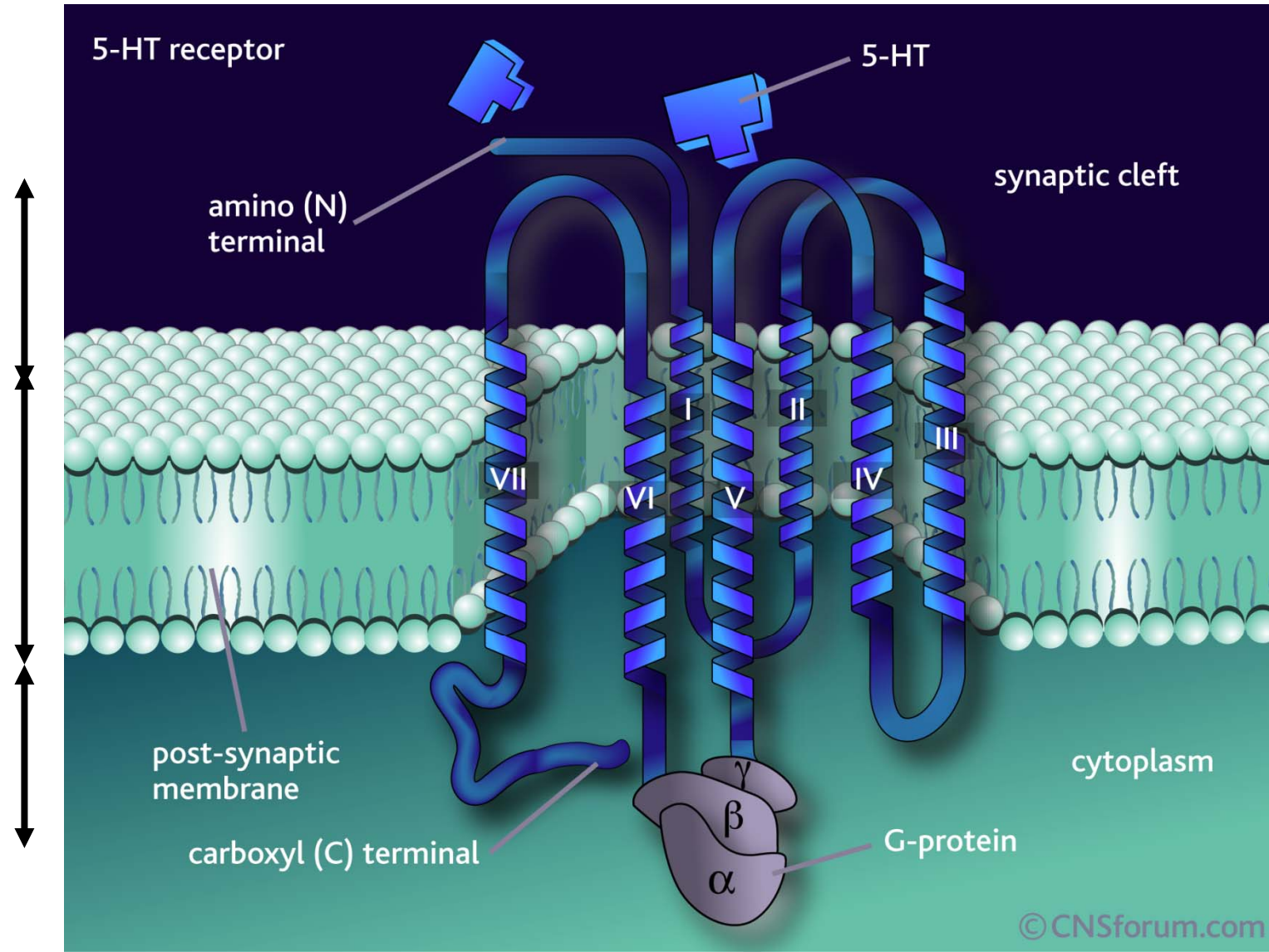
- Amino acids that make up the protein sequence have biochemical properties that influence their relative replaceability in an evolutionary scenario.



Transmembrane protein

Hydrophilic amino acids

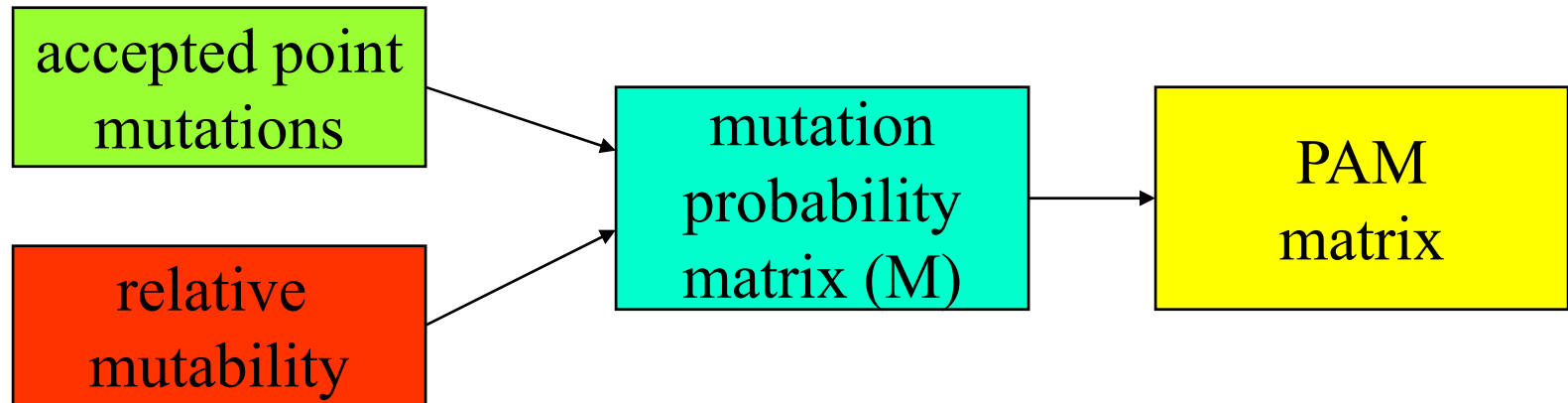
Hydrophobic amino acids



PAM matrices

- The factors that influence the probability of mutual substitution are numerous and various
- Therefore, direct observation of actual substitution rates is often the best way of deriving similarity scores for pairs of residues.
- A procedure towards this goal is based on the family of **PAM (Point Accepted Mutations)** scoring matrices.

PAM matrices



Accepted mutation

mutation that occurred and positively selected by the environment (did not cause the demise of the organism)

Relative mutability

measure of how often the amino acid changes

Mutation probability matrix (M):

probability that amino acid A is replaced with B after a given evolutionary interval.

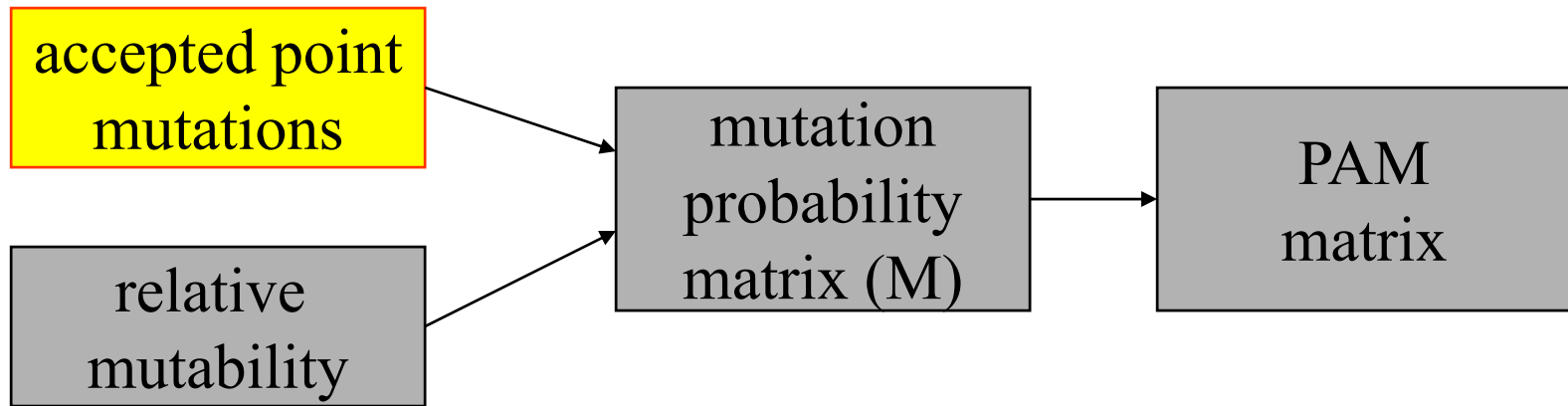
Amino Acid frequencies in the sequence data

| | <u>1978</u> | <u>1991</u> |
|-----|-------------|-------------|
| Leu | 0.085 | 0.091 |
| Ala | 0.087 | 0.077 |
| Gly | 0.089 | 0.074 |
| Ser | 0.070 | 0.069 |
| Val | 0.065 | 0.066 |
| Glu | 0.050 | 0.062 |
| Thr | 0.058 | 0.059 |
| Lys | 0.081 | 0.059 |
| Ile | 0.037 | 0.053 |
| Asp | 0.047 | 0.052 |
| Arg | 0.041 | 0.051 |
| Pro | 0.051 | 0.051 |
| Asn | 0.040 | 0.043 |
| Gln | 0.038 | 0.041 |
| Phe | 0.040 | 0.040 |
| Tyr | 0.030 | 0.032 |
| Met | 0.015 | 0.024 |
| His | 0.034 | 0.023 |
| Cys | 0.033 | 0.020 |
| Trp | 0.010 | 0.014 |

$$p_a = \frac{\text{observation of amino acid 'a'}}{\text{observation of any amino acid}} \quad \sum p_a = 1$$

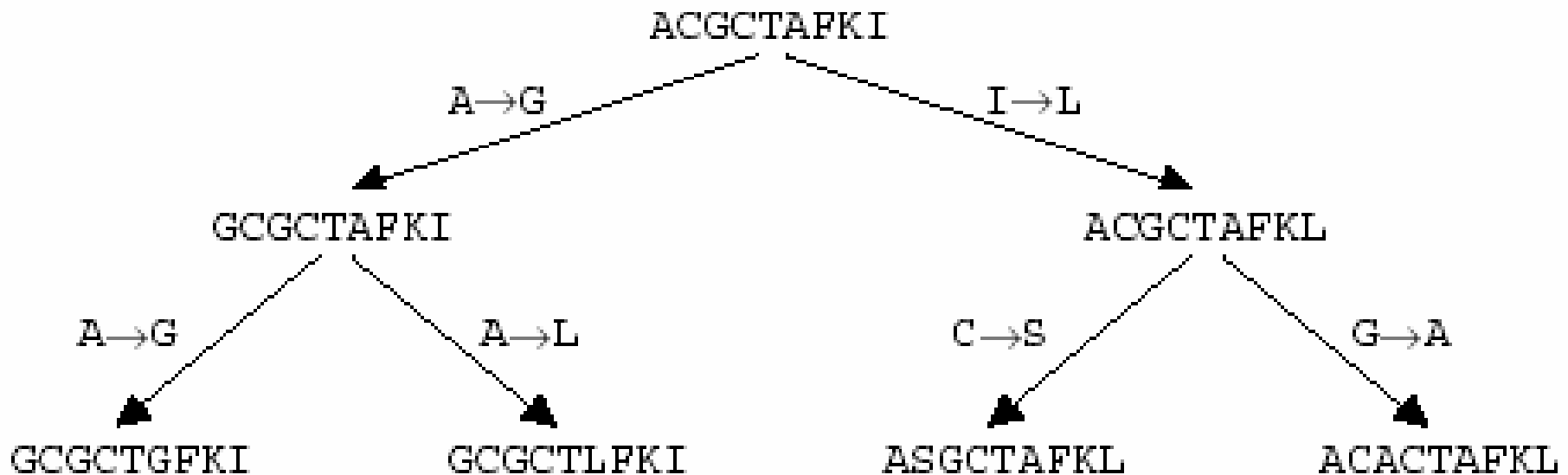
Accepted mutation

mutation that occurred and positively selected by the environment (did not cause the demise of the organism)



Accepted Point Mutations

- To identify accepted point mutations:
 - determine phylogenetic tree
 - compare mutations including the ancestral sequences (do not compare the sequences of end nodes directly)
 - To be able to do this, this tree was based on sequences with more than 85% identity (highly homologous sequences)

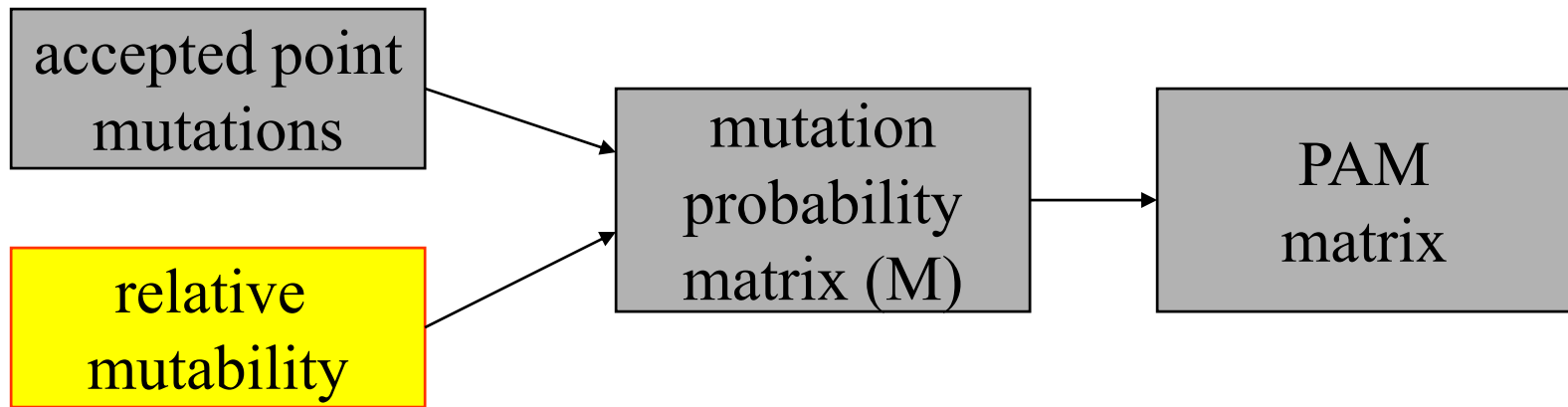


Accepted Point Mutations

- Accepted point mutation $f(ab)$ is accepted replacement of amino acid a with b
- Dayhoff et al used 71 groups of closely related proteins in 71 evolutionary trees in which they observed 1572 changes

Dayhoff, M.O., Eck, R.V., and Park, C.M., *in* Atlas of Protein Sequence and Structure 1972, Vol.5, ed. Dayhoff, M.O., pp.89-99, Nat. Biomed. Res. Found., Washington, D.C., 1972

- For each of the observed and inferred sequences in the tree, amino acid pair exchanges were tabulated in 20x20 matrix.
- It is assumed that probability of replacing a with b is equal to replacing b with a (symmetrical matrix)



Relative mutability

measure of how often the amino acid changes

Relative mutability of amino acids

- Relative mutability $m(a)$ of amino acid a is likelihood that 'a' is involved in mutation

Fraction of mutations in which amino acid 'a' is involved.

$$\frac{f_a}{f}$$

- $f(a)$ is number of mutations in amino acid 'a'
- f is total number of mutations

$$100p_a$$

Number of a's in sequence of 100 residues

$$m_a = \frac{\frac{f_a}{f}}{100p_a}$$

Fraction of 'a' mutations per $100p_a$ a's

$$m_a = \frac{f_a}{100fp_a}$$

Example

$$\frac{f_a}{f} = 0.5$$

$$100p_a = 100 * 0.05 = 5$$

$$\frac{f_a}{f} = 0.5$$

$$100p_a = 100 * 0.1 = 10$$

$$m_a = \frac{0.5}{5} = 0.1$$



If 'a' is abundant then relative mutability is lower (for constant fraction)

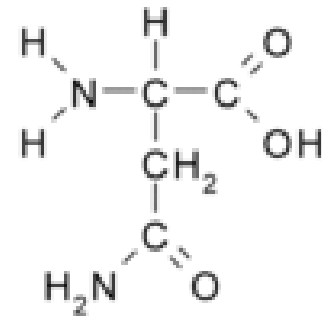
$$m_a = \frac{0.5}{10} = 0.05$$

Relative mutability of amino acids

| | 1978 | 1991 |
|-----|------|------|
| Ala | 100 | 100 |
| Cys | 20 | 44 |
| Asp | 106 | 86 |
| Glu | 102 | 77 |
| Peh | 41 | 51 |
| Gly | 49 | 50 |
| His | 66 | 91 |
| Ile | 96 | 103 |
| Lys | 56 | 72 |
| Leu | 40 | 54 |
| Met | 94 | 93 |
| Asn | 134 | 104 |
| Pro | 56 | 58 |
| Gln | 93 | 84 |
| Arg | 65 | 83 |
| Ser | 120 | 117 |
| Thr | 97 | 107 |
| Val | 74 | 98 |
| Trp | 18 | 25 |
| Tyr | 41 | 50 |

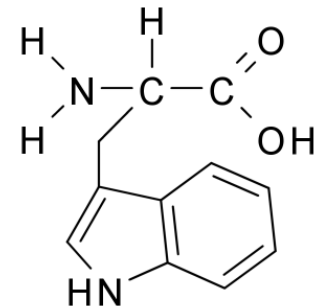
normalised to

Ala=100

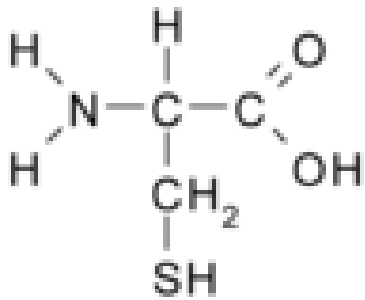


Asn=Asparagine

Trp=Tryptophan

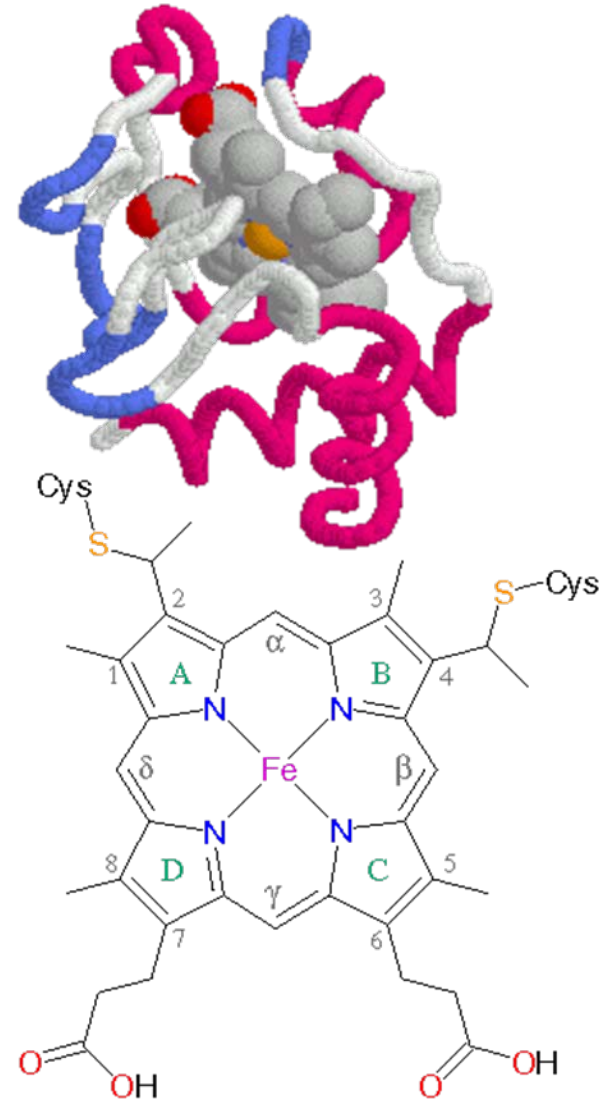


Relative mutability of amino acids



| | 1978 | 1991 |
|-----|------|------|
| Ala | 100 | 100 |
| Cys | 20 | 44 |
| Asp | 106 | 86 |
| Glu | 102 | 77 |
| Peh | 41 | 51 |
| Gly | 49 | 50 |
| His | 66 | 91 |
| Ile | 96 | 103 |

Cytochrome c



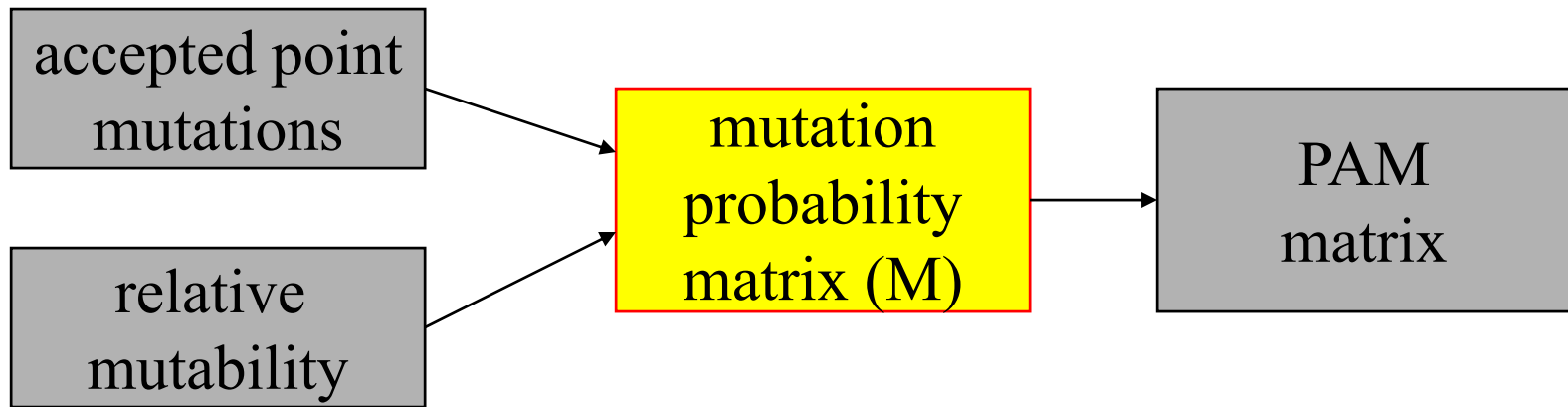
The heme of the cytochrome c

The immutability of cysteine is understandable.

Cysteine is known to have several unique, indispensable functions. It is the attachment site of heme groups in cytochrome and of FeS clusters in ferredoxin. It forms cross-links in other proteins such as chymotrypsin or ribonuclease.

It seldom occurs without having an important function.

| | | |
|-----|----|----|
| Trp | 18 | 25 |
| Tyr | 41 | 50 |



Mutation probability matrix (M):

probability that amino acid A is replaced with B after a given evolutionary interval.

Mutation probability matrix (M) 1

- Now we can calculate the probability that amino acid 'a' is replaced by 'b'
- This is conditional probability that *a* will change to *b* given that *a* changed, times the probability of *a* changing:

$$\begin{aligned}M_{ab} &= P(a \rightarrow b) \\ &= P(a \rightarrow b \mid a, t = 1) \cdot P(a \text{ changed}) \\ &= P(a \rightarrow b \mid a, t = 1) \cdot m_a\end{aligned}$$

number of amino acids 'a'
which are mutated in b

$$= \frac{f_{ab}}{f_a} m_a$$

number of amino acids 'a'
which are mutated

Mutation probability matrix 2

- We want to set $t=1$ when the expected number of mutations is 1% (0.01)

$$\sum_a \sum_{b \neq a} p_a M_{ab} = \text{weighted average}$$

$$= \sum_a p_a \left(\sum_{b \neq a} M_{ab} \right)$$

$$\sum_a p_a \left(\sum_{b \neq a} \frac{f_{ab}}{100 f p_a} \right)$$

$$\text{let } \sigma = \frac{1}{100} = 0.01$$

$$= \sum_a p_a \left(\sigma \sum_{b \neq a} \frac{f_{ab}}{f p_a} \right)$$

$$\text{note } \sum_{b \neq a} f_{ab} = f_a$$

$$= \sigma \sum_a p_a \frac{f_a}{f p_a}$$

$$= \sigma \quad \text{Thus } \sigma=0.01 \text{ represents overall change of 1\%}$$

Mutation probability matrix 3

- Next, determine the probabilities of **no** mutation M_{aa}

$$\sum_{b \neq a} (M_{ab}) + M_{aa} = 1$$

$$\sum_{b \neq a} \left(\sigma \frac{f_{ab}}{f_a} m_a \right) + M_{aa} = 1 \quad (\text{note that } \sum_{b \neq a} f_{ab} = f_a)$$

$$M_{aa} = 1 - \sigma m_a$$

Mutation probability matrix 4

- M has the following properties:

$$\sum_b M_{ab} = 1$$

M represent real probabilities

$$\sum_a p_a M_{aa} = 0.99$$

Probability of change=0.01
1 of every 100 amino acids

The amount of evolution that will change 1 in 100 amino acids on average is referred to as 1 PAM evolutionary distance

Mutation probability matrix 5 (assymmetric)

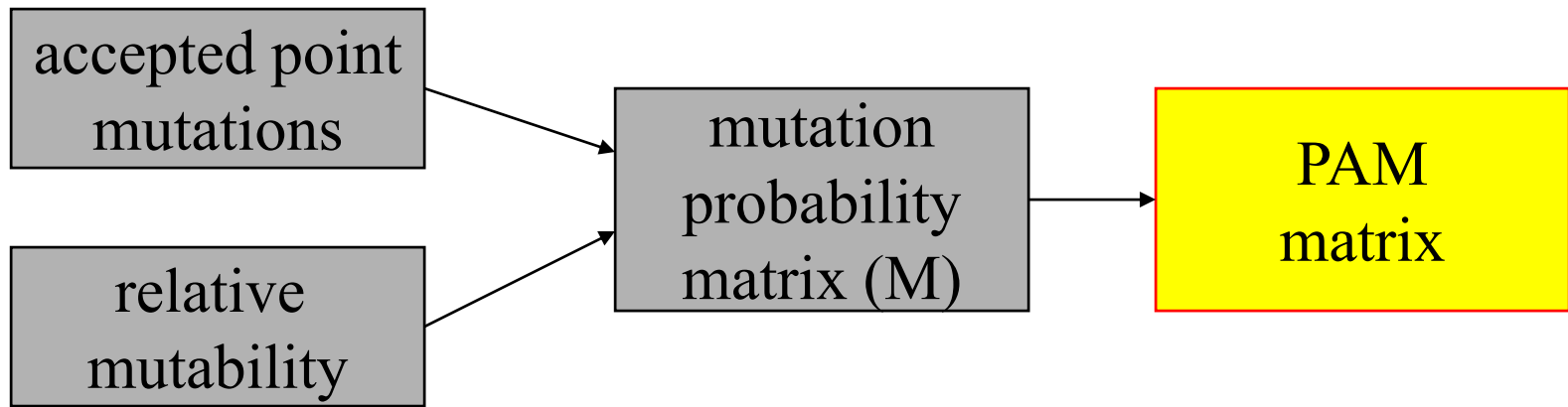
ORIGINAL AMINO ACID

| | A | R | N | D | C | Q | E | G | H | I | L | K | M | F | P | S | T | W | Y | V |
|-------|-------|-------|-------|-------|-------|-------|-------|-------|------|-------|-------|-------|-------|-------|-------|-------|-------|------|------|------|
| | Ala | Arg | Asn | Asp | Cys | Gln | Glu | Gly | His | Ile | Leu | Lys | Met | Phe | Pro | Ser | Thr | Trp | Tyr | Val |
| A Ala | 99.67 | 2 | 9 | 10 | 3 | 8 | 17 | 21 | 2 | 6 | 4 | 2 | 6 | 2 | 22 | 35 | 32 | 0 | 2 | 18 |
| R Arg | 1 | 99.13 | 1 | 0 | 1 | 10 | 0 | 0 | 10 | 3 | 1 | 19 | 4 | 1 | 4 | 6 | 1 | 8 | 0 | 1 |
| N Asn | 4 | 1 | 99.22 | 36 | 0 | 4 | 6 | 6 | 21 | 3 | 1 | 13 | 0 | 1 | 2 | 20 | 9 | 1 | 4 | 1 |
| D Asp | 6 | 0 | 42 | 99.59 | 0 | 6 | 53 | 6 | 4 | 1 | 0 | 3 | 0 | 0 | 1 | 5 | 3 | 0 | 0 | 1 |
| C Cys | 1 | 1 | 0 | 0 | 99.73 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 5 | 1 | 0 | 3 | 2 |
| Q Gln | 3 | 9 | 4 | 5 | 0 | 99.76 | 27 | 1 | 23 | 1 | 3 | 6 | 4 | 0 | 6 | 2 | 2 | 0 | 0 | 1 |
| E Glu | 10 | 0 | 7 | 56 | 0 | 35 | 99.55 | 4 | 2 | 3 | 1 | 4 | 1 | 0 | 3 | 4 | 2 | 0 | 1 | 2 |
| G Gly | 21 | 1 | 12 | 11 | 1 | 3 | 7 | 99.95 | 1 | 0 | 1 | 2 | 1 | 1 | 3 | 21 | 3 | 0 | 0 | 5 |
| H His | 1 | 2 | 18 | 3 | 1 | 20 | 1 | 0 | 99.2 | 0 | 1 | 1 | 0 | 2 | 3 | 1 | 1 | 1 | 4 | 1 |
| I Ile | 2 | 2 | 3 | 1 | 2 | 1 | 2 | 0 | 0 | 99.72 | 9 | 2 | 12 | 7 | 0 | 1 | 7 | 0 | 1 | 33 |
| L Leu | 3 | 1 | 3 | 0 | 0 | 6 | 1 | 1 | 4 | 22 | 99.17 | 2 | 45 | 13 | 3 | 1 | 3 | 4 | 2 | 15 |
| K Lys | 2 | 37 | 25 | 6 | 0 | 12 | 7 | 2 | 2 | 4 | 1 | 99.26 | 20 | 0 | 3 | 8 | 11 | 0 | 1 | 1 |
| M Met | 1 | 1 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 5 | 8 | 4 | 99.74 | 1 | 0 | 1 | 2 | 0 | 0 | 4 |
| F Phe | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 2 | 8 | 6 | 0 | 4 | 99.16 | 0 | 2 | 1 | 3 | 28 | 0 |
| P Pro | 13 | 5 | 2 | 1 | 1 | 8 | 3 | 2 | 5 | 1 | 2 | 2 | 1 | 1 | 99.26 | 12 | 4 | 0 | 0 | 2 |
| S Ser | 28 | 11 | 34 | 7 | 11 | 4 | 6 | 16 | 2 | 2 | 1 | 7 | 4 | 3 | 17 | 98.40 | 38 | 5 | 2 | 2 |
| T Thr | 22 | 2 | 13 | 4 | 1 | 3 | 2 | 2 | 1 | 11 | 2 | 8 | 6 | 1 | 5 | 32 | 98.71 | 0 | 2 | 9 |
| W Trp | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 99.5 | 1 | 0 |
| Y Tyr | 1 | 0 | 3 | 0 | 3 | 0 | 1 | 0 | 4 | 1 | 1 | 0 | 0 | 21 | 0 | 1 | 1 | 2 | 99.5 | 1 |
| V Val | 13 | 2 | 1 | 1 | 3 | 2 | 2 | 3 | 3 | 57 | 11 | 1 | 17 | 1 | 3 | 2 | 10 | 0 | 2 | 99.0 |

Expected value
upper triangle = 0.05

Expected value
lower triangle = 0.05

Expected value
over diagonal = 0.99



The PAM score matrix

To score alignment (or segment): multiply probabilities:

$$s = \prod \frac{M_{ab}}{p_b}$$

The scores are logarithms:

we can **sum the individual scores** for each pair of residues

$$s = \prod \frac{M_{ab}}{p_b} \Rightarrow \log(s) = \log\left(\prod \frac{M_{ab}}{p_b}\right) \Rightarrow \log(s) = \log\left(\frac{M_{ab}}{p_b}\right) + \log\left(\frac{M_{ab}}{p_b}\right) + \dots\dots\dots$$

The PAM score matrix

Scores for specific alignment position is calculated as follows:

$$s_{ab} = Q \log_{10} \frac{M_{ab}}{p_b}$$

Q=10 to reduce discrepancy
between correct value and integer
approximation

The PAM score matrix

$$s_{ab} = Q \log_{10} \frac{M_{ab}}{p_b}$$

This can be rewritten as:

$$s_{ab} = Q \log \frac{f_{ab}}{fp_a p_b} = \log \left(\frac{q_{ab}}{p_a p_b} \right)$$

score reflects
bias in observing
 $a \rightarrow b$ w.r.t. background
frequencies

q_{ab} are the target frequencies associated with the scores

PAM matrix

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

Sulfhydryl

small
hydrophilic

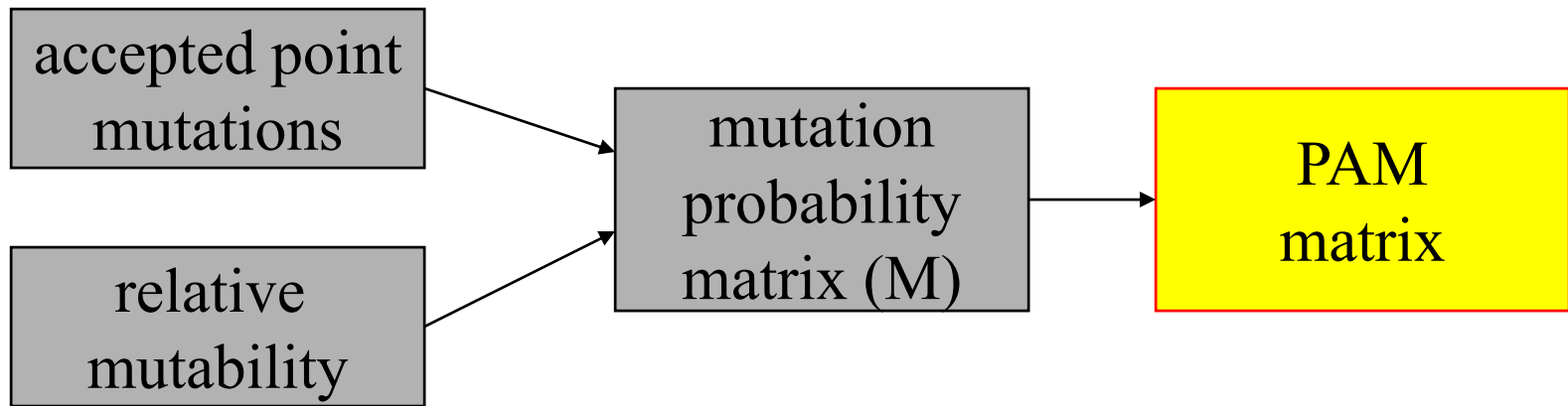
acid amide and
hydrophilic

basic

small
hydrophobic

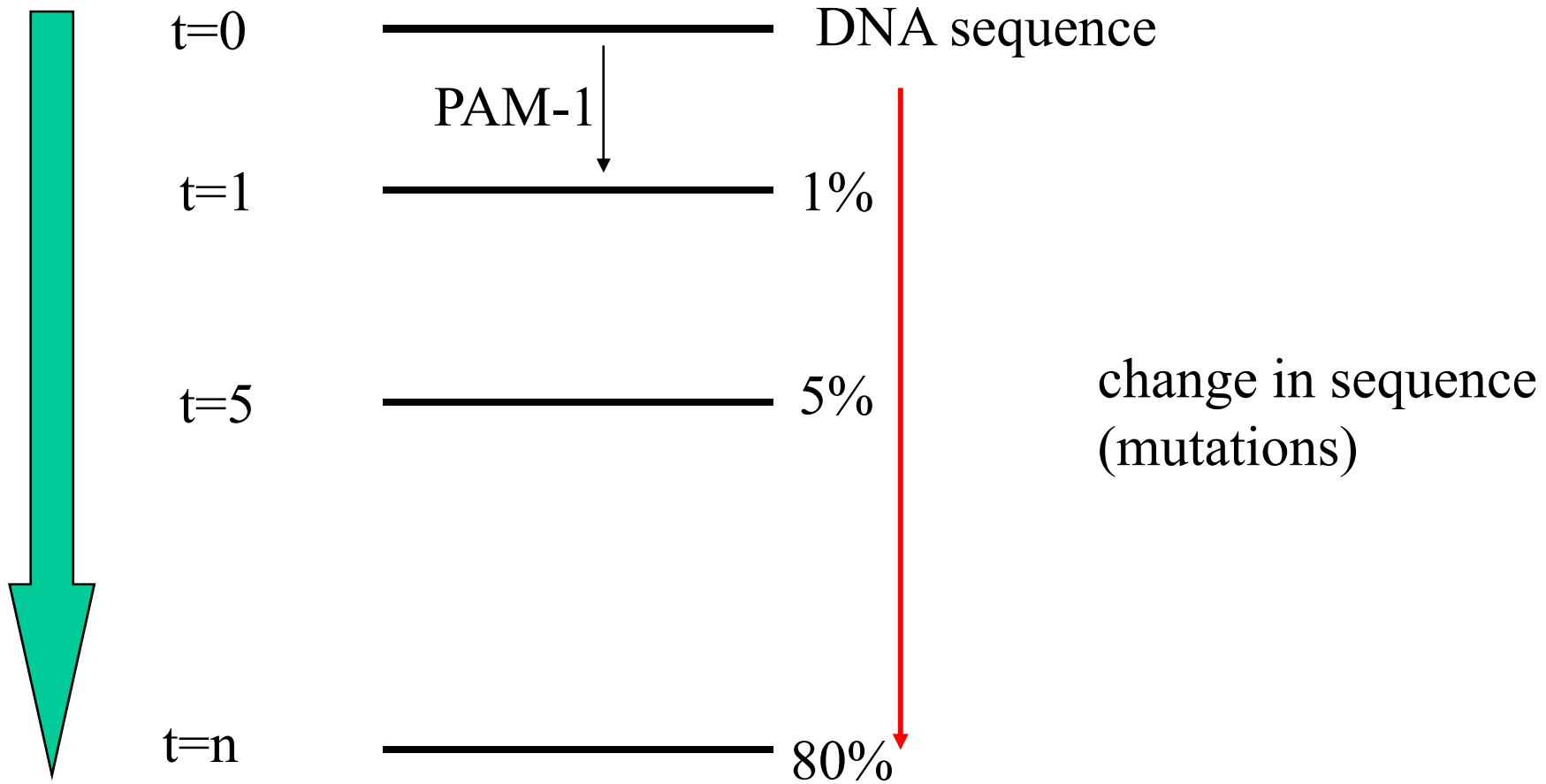
Aromatic

Positive scores for similar amino acids



Higher order PAM matrices

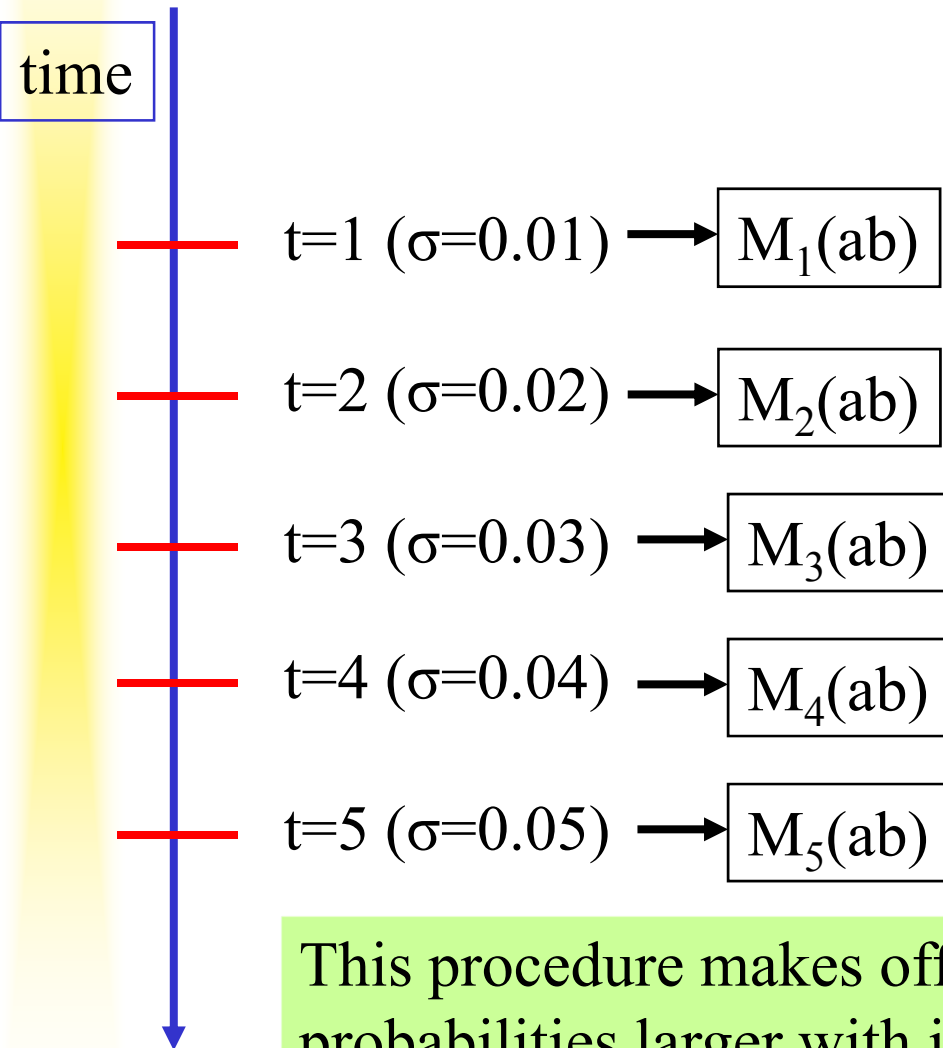
PAM and evolution



PAM Matrix

- The **PAM-1 matrix** was derived on groups of sequences with less than 15% differences
- We have scaled M_{ab} such that the expected amount of change reflected by this matrix is 1% (1 mutation per 100 amino acids)
- We could have chosen $\sigma > 0.01$, e.g. $\sigma = 0.05$ to obtain a PAM-5 matrix that reflects 5 mutations per 100 amino acids
- However, to derive valid PAM matrices in this way, the set of sequences must reflect the required amount of change.
- Simply choosing a higher σ would neglect 'overlapping' mutations.

Increasing amount of expected change



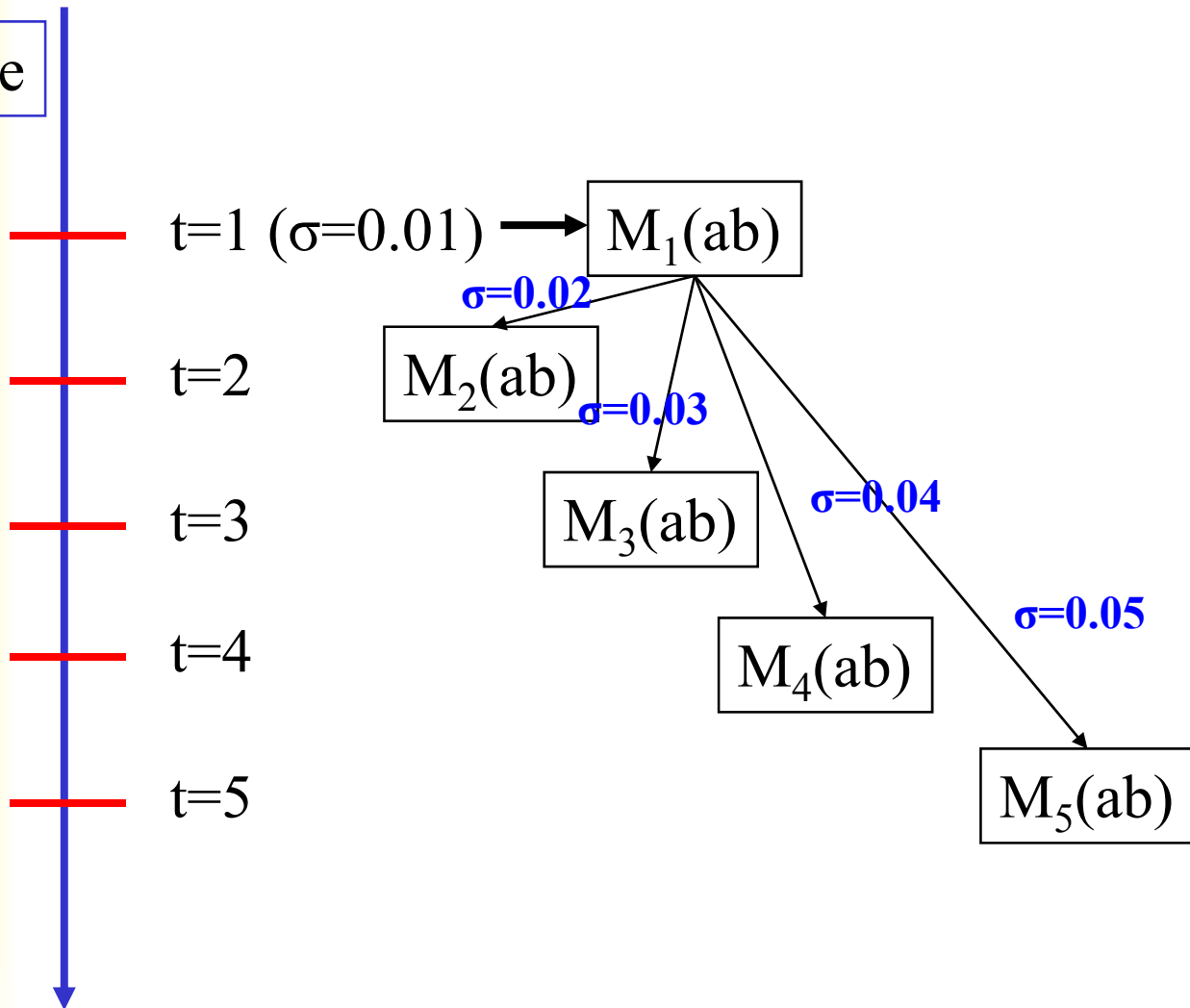
This procedure makes off-diagonal probabilities larger with increasing sigma.

This neglects back-substitutions → off-diagonal P are too large

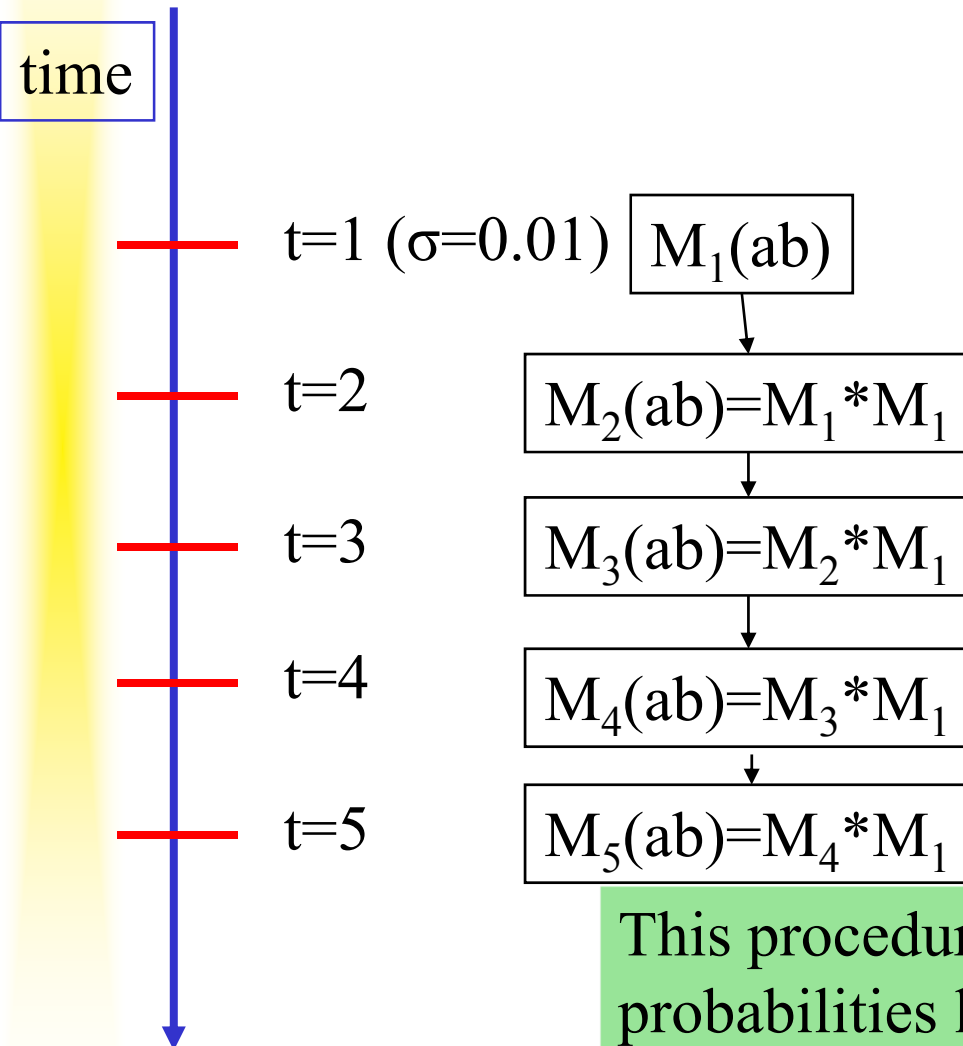
Higher order PAM matrices

- What is the probability that $a \rightarrow b$ in two PAM units of evolution ($\sigma=0.02$).
- Possible routes:
 - $A \rightarrow C \rightarrow B$
 - $A \rightarrow A \rightarrow B$
 - $A \rightarrow B \rightarrow B$
- Probability = $p(ac)p(cb)+p(aa)p(ab)+p(ab)p(bb)$
- This is exactly M_{ab}^2
- Thus PAM-N is directly obtained from M_{ab}^N

Increasing amount of expected change



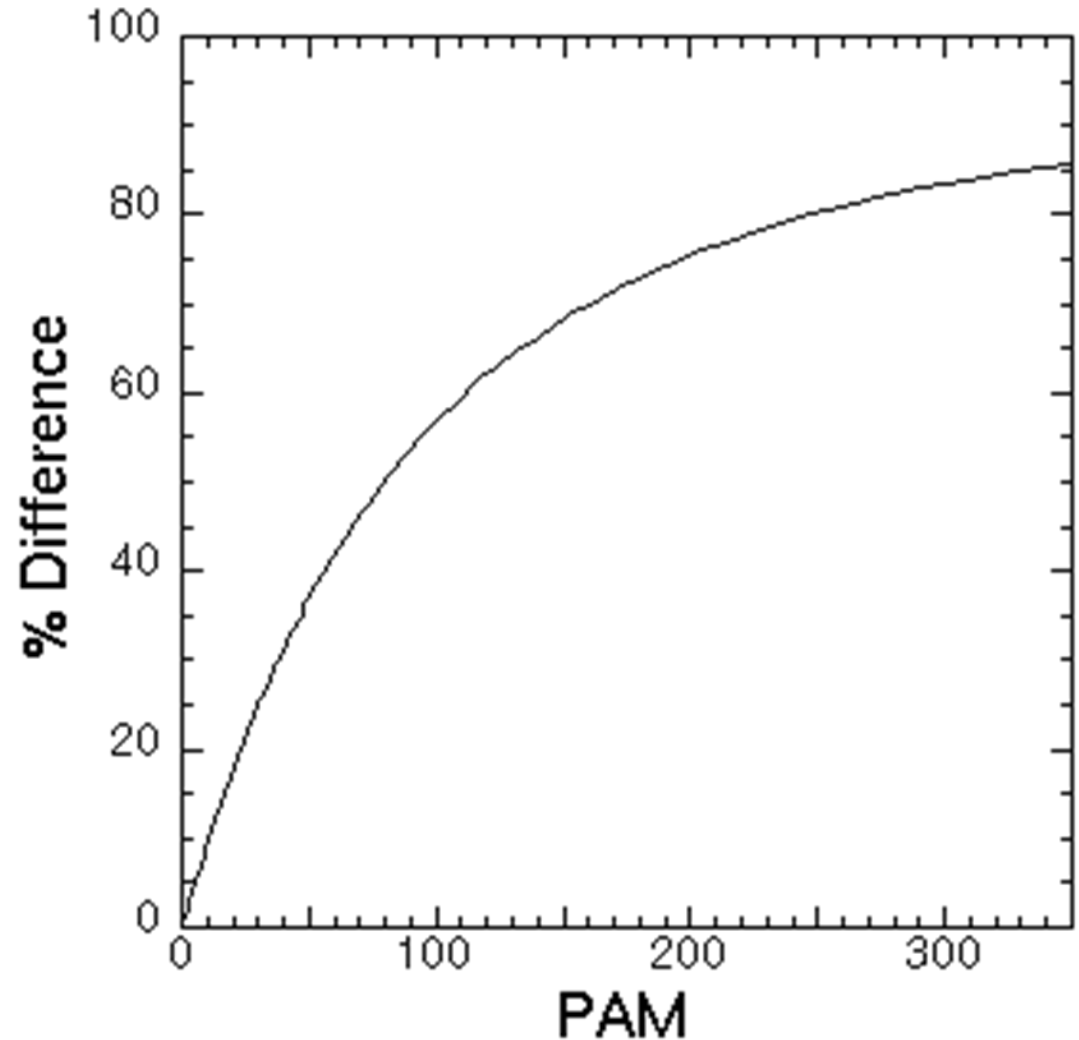
Increasing amount of expected change



This procedure makes off-diagonal probabilities larger with increasing multiplication. Does not neglect back-substitutions

Higher order PAM matrices: % Difference

| %Difference | PAM |
|-------------|------------|
| 1 | 1 |
| 5 | 5 |
| 10 | 11 |
| 15 | 17 |
| 20 | 23 |
| 25 | 30 |
| 30 | 38 |
| 35 | 47 |
| 40 | 56 |
| 45 | 67 |
| 50 | 80 |
| 55 | 94 |
| 60 | 112 |
| 65 | 133 |
| 70 | 159 |
| 75 | 195 |
| 80 | 246 |
| 85 | 328 |



Higher order PAM matrices: % Difference

Consider alignment with few identities:

```
..... C S T P T A H R K .....  
                S                K  
..... A S P A K H R T K.....
```

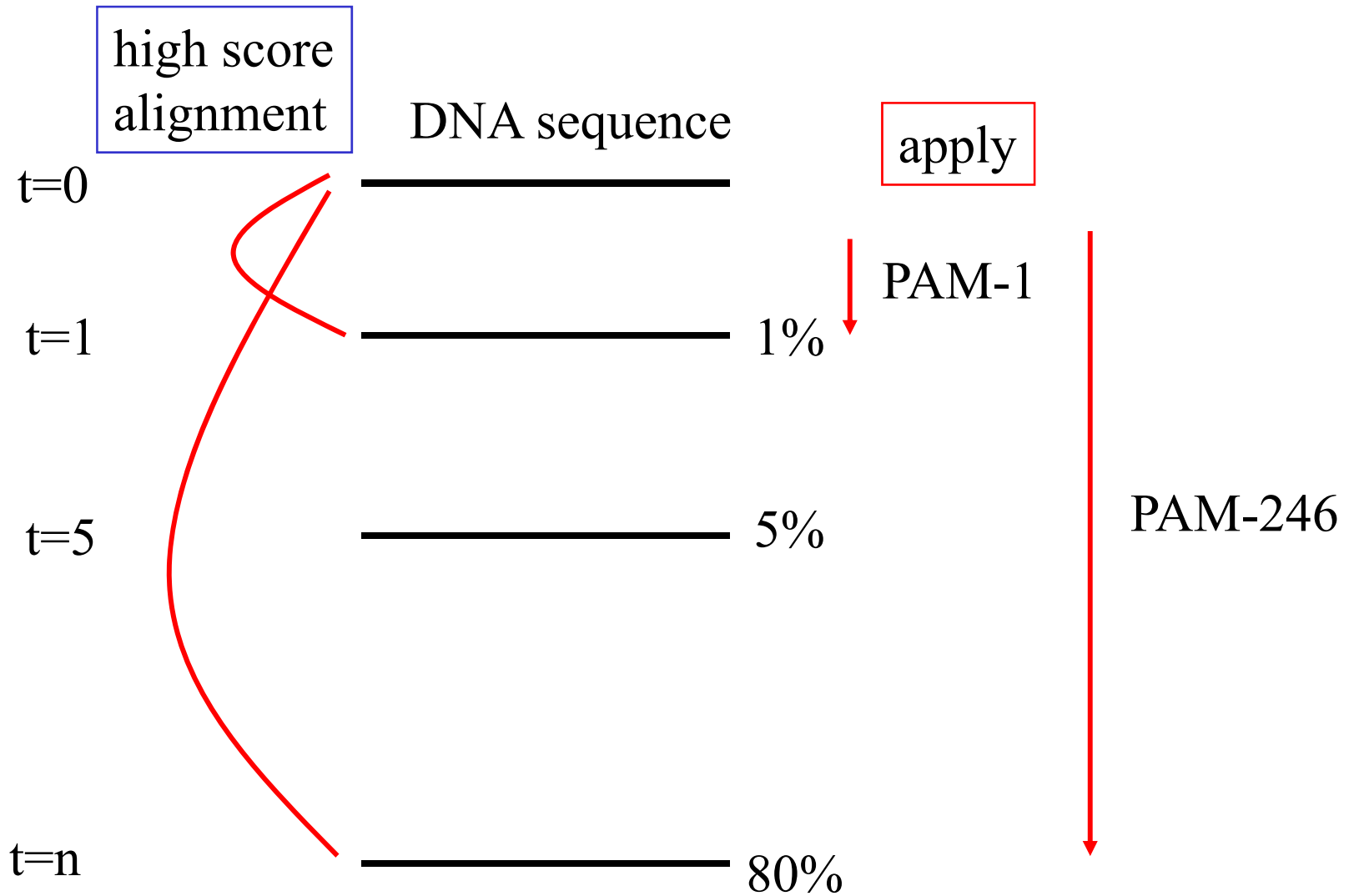
| | | | | | | | | | | |
|---------|-----|---|----|----|----|-----|-----|-----|---|-------|
| PAM-10 | -10 | 7 | -7 | -4 | -6 | -11 | -10 | -19 | 7 | (-50) |
| PAM-500 | -2 | 1 | 1 | 1 | 0 | 0 | -2 | -6 | 4 | (-3) |

Some transitions are more likely at larger distances

Some transitions are less likely at larger distances

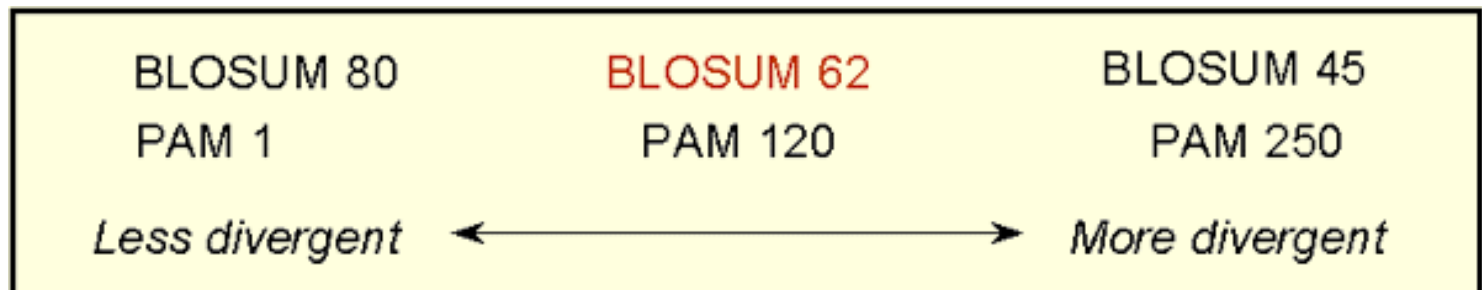
Choice of scoring matrix critical for detecting similarity

PAM and evolution



BLOSUM matrices

- PAM-k matrices are derived from PAM-1 matrix by extrapolation
- **Better approach:** get measure of differences between two proteins that are distantly related.
- Based on the BLOCKS database: direct observation of sequences; no evolutionary model
- BLOSUM: **BLOcks SUBstitution Matrix**
- BLOSUM62 matrix is calculated from protein blocks in which two sequences are more than 62% identical.



Score assignments

Score assignments

To identify interesting patterns in individual or sets of sequences we must assign scoring values to residues.

Scoring assignments for single nucleotide or amino-acid sequences can be provided by:

- Biochemical properties (charge, hydrophobicity)
- Physical properties (molecular weight, shape)
- Associations with secondary structure (alpha helices, beta-strands, turns, open coils)

For sequence alignments scores reflect nucleotide or amino acid similarity.

Natural Scoring Assignments

Scores based on charge

- Lysine, arginine $S=+1$
- Aspartate, Glutamate $S=-1$
- Histidine $S=0.04$ (pH=7.2 in blood serum)
 $S=0.44$ (pH=6.1 in muscle cells)
- other $S=0$

Scores associated with a run of a particular letter type

- Score of letter A is +1, score of other letters = $-\infty$

If we have a sequence

XXXXXXXXXXXX**AAAAAAAA**XXXXXXXXXXXXXXXX

only the run of eight A's has a positive score.

In fact it has the highest possible score ($S=8$)

Natural Scoring Assignments

Scores derived from target frequencies

In a random sequence the letters are sampled with probabilities $\{p_1, p_2, \dots, p_r\}$.

Let $\{q_1, q_2, \dots, q_r\}$ be a set of target frequencies.

Then scores are log-likelihood ratios:

$$s_i = \log \left(\frac{q_i}{p_i} \right)$$

$$s_{ab} = \log_2 \left(\frac{q_{ab}}{p_a p_b} \right)$$

This is used for
the PAM matrices

Example

Target frequencies (q) were derived from set of selected example sequences

| | p | q | p/q | log ₁₀ (q/p) |
|---|------|------|-----|-------------------------|
| A | 0,25 | 0,05 | 0,2 | -0,70 |
| T | 0,25 | 0,05 | 0,2 | -0,70 |
| C | 0,25 | 0,05 | 0,2 | -0,70 |
| G | 0,25 | 0,85 | 3,4 | 0,53 |

Sequences to 'test'

catgaaaaa s1

catgggggg s2>>s1

Restrictions on set of scores

1. At least one score is positive
2. The expected score $E = \sum p_i s_i < 0$

If all scores would be negative then the maximum segment would always have a length of one (increasing the segment would result in a more negative score).

The second restriction ensures that the maximal segment is not the complete sequence (increasing the sequence would on average increase the score).

Determination of statistical significance of a score

In the remaining we assume that we have selected a scoring scheme (e.g., a PAM matrix).

Intermezzo: logarithms

$$\log(X) = A \Rightarrow X = 10^A$$

$$\log(100) = 2 \Rightarrow 100 = 10^2$$

$$\ln(X) = B \Rightarrow X = e^B$$

$$\ln(2.718) = 1 \Rightarrow e^1 = 2.718$$

Base of the logarithm

Base is 10

$$\log(X) \Leftrightarrow \log_{10}(X)$$

$$\ln(X) \Leftrightarrow \log_e(X)$$

Base is $e = 2.718$

Logarithm with different bases

$$\log_2(X) = A \Leftrightarrow X = 2^A$$

$$\log_2(4) = 2 \Leftrightarrow 4 = 2^2$$

Base is 2
Unit = bits

$$\log_p(X) = A \Leftrightarrow X = p^A$$

Base is p
(e.g. probability)

$$\log_{1/p}(X) = A \Leftrightarrow X = \left(\frac{1}{p}\right)^A$$

Conversion of logarithms with different base

$$\log_a(X) = \frac{\log_{10}(X)}{\log_{10}(a)} = \frac{\ln(X)}{\ln(a)}$$

$$\log_2(X) = \frac{\log_{10}(X)}{\log_{10}(2)} = \frac{\ln(X)}{\ln(2)}$$

$$\log_e(X) = \frac{\log_{10}(X)}{\log_{10}(e)} = \frac{\ln(X)}{\ln(e)} = \ln(X)$$

Conversion of logarithms with different base

$$\begin{aligned}\log_{1/p}(X) &= \frac{\log_{10}(X)}{\log_{10}(1/p)} \\ &= \frac{\log(X)}{\log(1) - \log(p)} \\ &= -\frac{\log(X)}{\log(p)}\end{aligned}$$

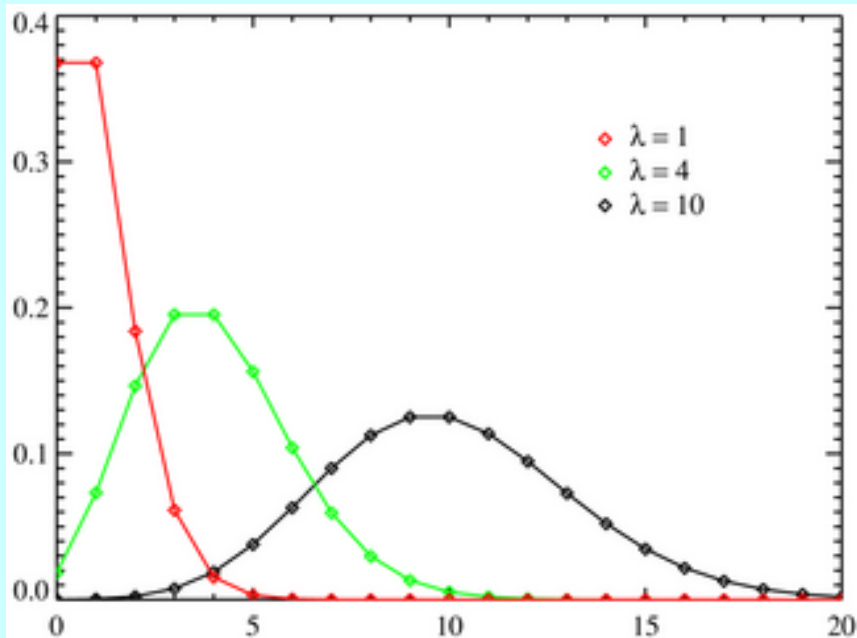
Note: $\log(1)=0$

Intermezzo: Poisson distribution

$$P_n(\mu) = \frac{e^{-\mu} \mu^n}{n!}$$

$P(\mu)$ probability of getting n counts ($=0, 1, 2, \dots$)

μ average of distribution



variance == mean

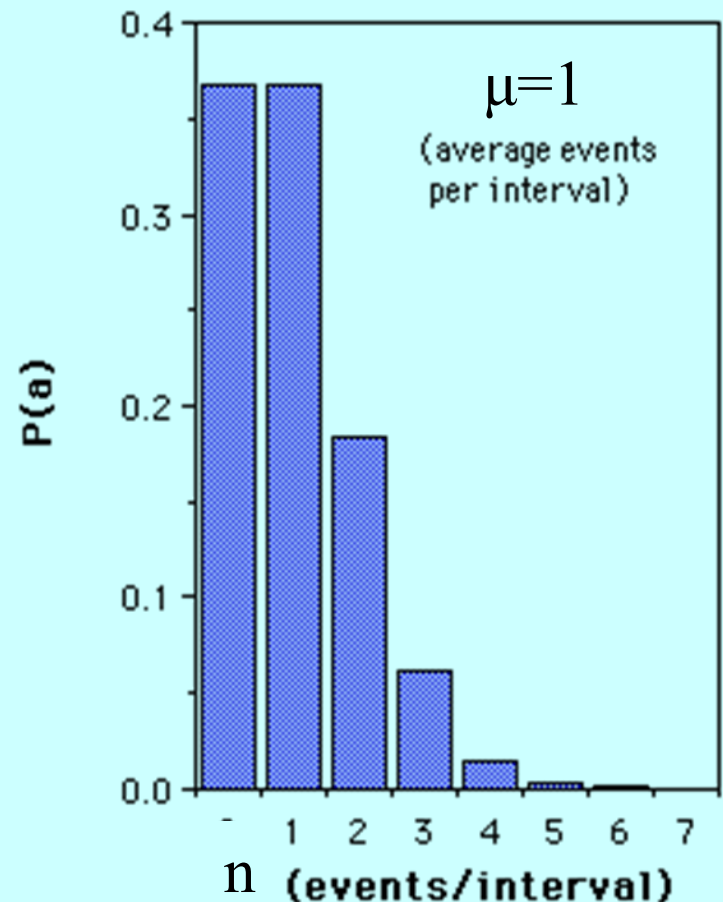
Intermezzo: Poisson distribution



Randomly placed dots over 50 scale divisions.
On average $\mu=1$ dot per interval

$$P_n(\mu) = \frac{e^{-\mu} \mu^n}{n!}$$

$P(\mu)$ probability of getting n counts
 μ average of distribution



Poisson distribution: Example 1

$$P_n(\mu) = \frac{e^{-\mu} \mu^n}{n!}$$

$P_n(\mu)$ probability of getting n counts
 μ average of distribution

Average number of phone calls in 1 hour = 2.1

What is probability of getting 4 calls?

Answer

$$P_{n=4}(\mu = 2.1) = \frac{e^{-2.1} 2.1^4}{4!} = 0.0992$$

Poisson distribution: Example 2

$$P_n(\mu) = \frac{e^{-\mu} \mu^n}{n!}$$

$P_n(\mu)$ probability of getting a discrete value n
 μ average of distribution

Average number of phone calls in 1 hour = 2.1

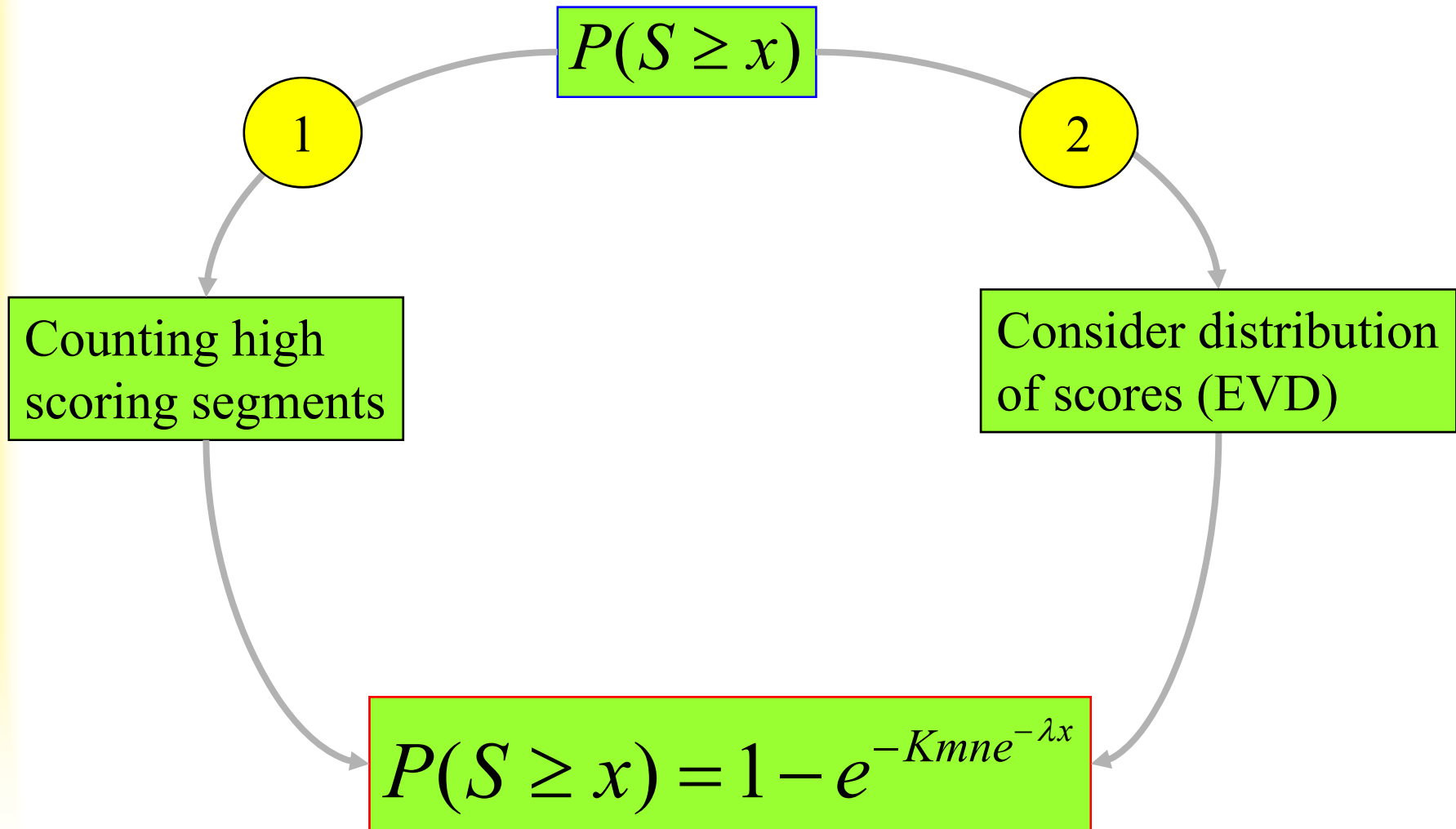
What is probability of getting 0 calls?

Answer

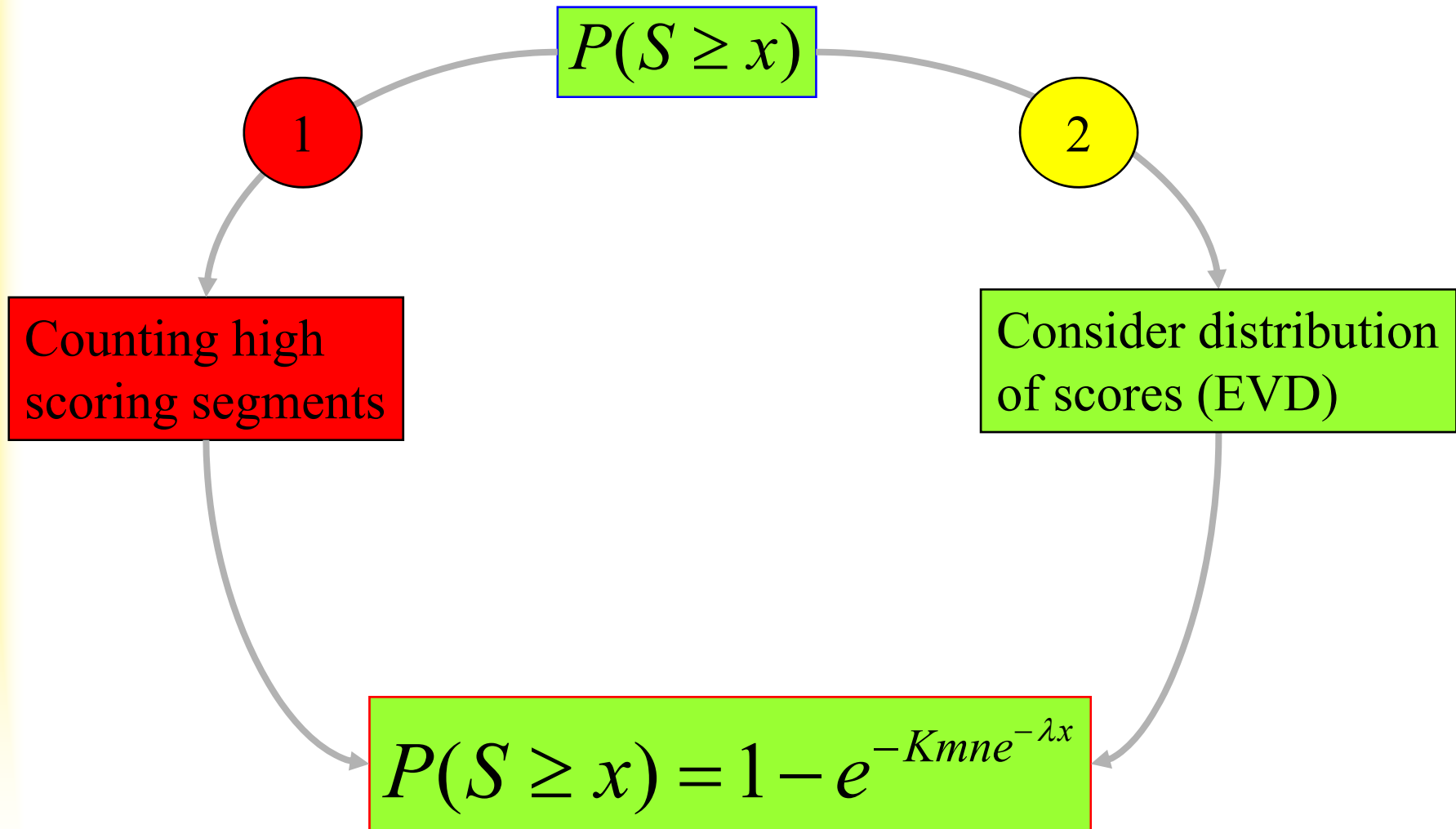
$$P_{n=0}(\mu = 2.1) = \frac{e^{-2.1} 2.1^0}{0!} = e^{-2.1} = 0.122$$

$$P_{n=0}(\mu = 0) = \frac{e^{-x} 0^n}{0!} = e^{-x}$$

Calculating statistical significance of scores



Calculating statistical significance of scores



Random model

‘random’ model

- random nucleotide or protein sequences
- provides a benchmark for analyzing various data statistics.

A random sequence: the coin model

- Create random DNA sequences by flipping a coin
 - Head (H) → region of interest
 - Tail (T)

For example: TTT**HHHHHHH**TTTTTT

Corresponds to an sequence with e.g.
transmembrane region

Question: what is the probability of finding a stretch of n heads?

The coin

The coin has

*probability p of scoring a head (H)

*probability $q=1-p$ of scoring a tail (T).

If $p=0.5$ (normal coin).

Length of run of L heads

$\Pr(L) = p^L$ Probability of getting run of L heads

$E(L) = n * p^L$ Expected number of runs of length L

n = number of trials

Suppose $p=0.5$. What is the chance of getting 5 heads?

Answer $\Pr(L=5) = (0.5)^5 = 0.03125$

Suppose we toss 14 times. How many times do we observe run of 5 heads?



$14 - 4 = 10$ ($=n$) places to start
5 heads

Answer: $E(L=5, n=10) = 10 * 0.03125 = 0.3125$

Longest run of heads

Suppose we toss n times.

What is the **longest run** we will observe?

Answer: the longest run has the smallest probability to occur and is expected to occur only **once in n trials**

Denote longest run by $R_L \rightarrow E(R_L)=1$

$$E(L) = n * p^L \Rightarrow E(R_L) = n * p^{R_L} = 1$$

Example: Longest run of heads

Longest run of heads $R = \frac{1}{p} \log(n)$.

If $p=0.5$ (normal coin).

For $n=100$ trials $\rightarrow R=6.65$

..... TTT **HHHHHHH** TTT **HHH** TTT **H** T **HHHH** T **HH** T

A random alignment: the coin model

- Create two aligned random DNA sequences by flipping a coin
 - Head (H) → match in alignment
 - Tail (T) → mismatch in alignment

For example: TTT**HHHHHHH**TTTTTT

Corresponds to an alignment with 7 arbitrary matches:

```
CATGGATGACCGTGCC  
ATAGGATGACAAAAAAA
```

Longest random alignment

In alignment, sequences are shifted back and forth to find regions that can be aligned.

Alignment can start at $n \times m$ places

| | H | I | K | T | Q | S | N | A | I | L |
|---|---|---|---|---|---|---|---|---|---|---|
| H | ● | | | | | | | | | |
| E | | | | | | | | | | |
| S | | | | | | ● | | | | |
| R | | | | | | | | | | |
| A | | | | | | | | ● | | |
| I | | ● | | | | | | | ● | |
| Q | | | | | ● | | | | | |
| V | | | | | | | | | | |

Comparison of two protein sequences, with identities indicated as black circles. Assuming the residues were drawn from a population of 20, each with the same probability, the probability of an identical match is $p = 0.05$. In this example, there are $m = 10 \times n = 80$ boxes, so $E() = mnp = 80 \times 0.05 = 4$ matches are expected by chance. The probability of two successive matches is $p^2 = 1/(20)^2$ so a run of two matches is expected about $nmp^2 = 80 \times 1/(20)^2 = 0.2$ times by chance.

$$\text{Thus } R = 1/p \log(nm).$$

Longest random alignment

In alignment, sequences are shifted back and forth to find regions that can be aligned.

```
CATGGATGACCGTAAA  
  AGGATAGGATGACAAAAAA
```

```
CATGGATGACCGTAAA  
ATAGGATGACAAAAAA
```

```
                CGGAATGGATGACCGTAAA  
ATAGAGATGACAAAGGA
```

Therefore, $R = \frac{1}{p} \log(nm)$.

Longest random alignment

Therefore, $R = 1/p \log(nm)$.

Example:

4 nucleotides (C,A,T,G)

$p=0.25$ (every nucleotide has occurs 25%)

Chance of aligning identical nucleotides = $0.25*0.25$

However, 4 nucleotides, chance for match = $4*0.25*0.25$

$p=0.25$

$n=m=100$

→ $R=6.65$

Mean $E(M)$ of longest run (M)

100 alignments ($p=0.25$, n,m):

| | <u>Number of heads:</u> |
|---------------------------------|-------------------------|
| Exp 1: H H T T H H.....H T | $M=7$ |
| Exp 2: T H H H H TT T | $M=5$ |
| | |
| | |
| Exp 100: T T H H H H H H H | $M=8$ |

Average = $E(M) = 6.5$

The mean of longest runs that are expected to occur only **once** in every experiment

Mean of the longest run

More formal formula for **mean** of longest match M :

$$E(M) \cong \log_{1/p}(mn) + \log_{1/p}(1-p) + \gamma \log(e) - \frac{1}{2}$$

where gamma $\gamma=0.577$ (Euler's constant).

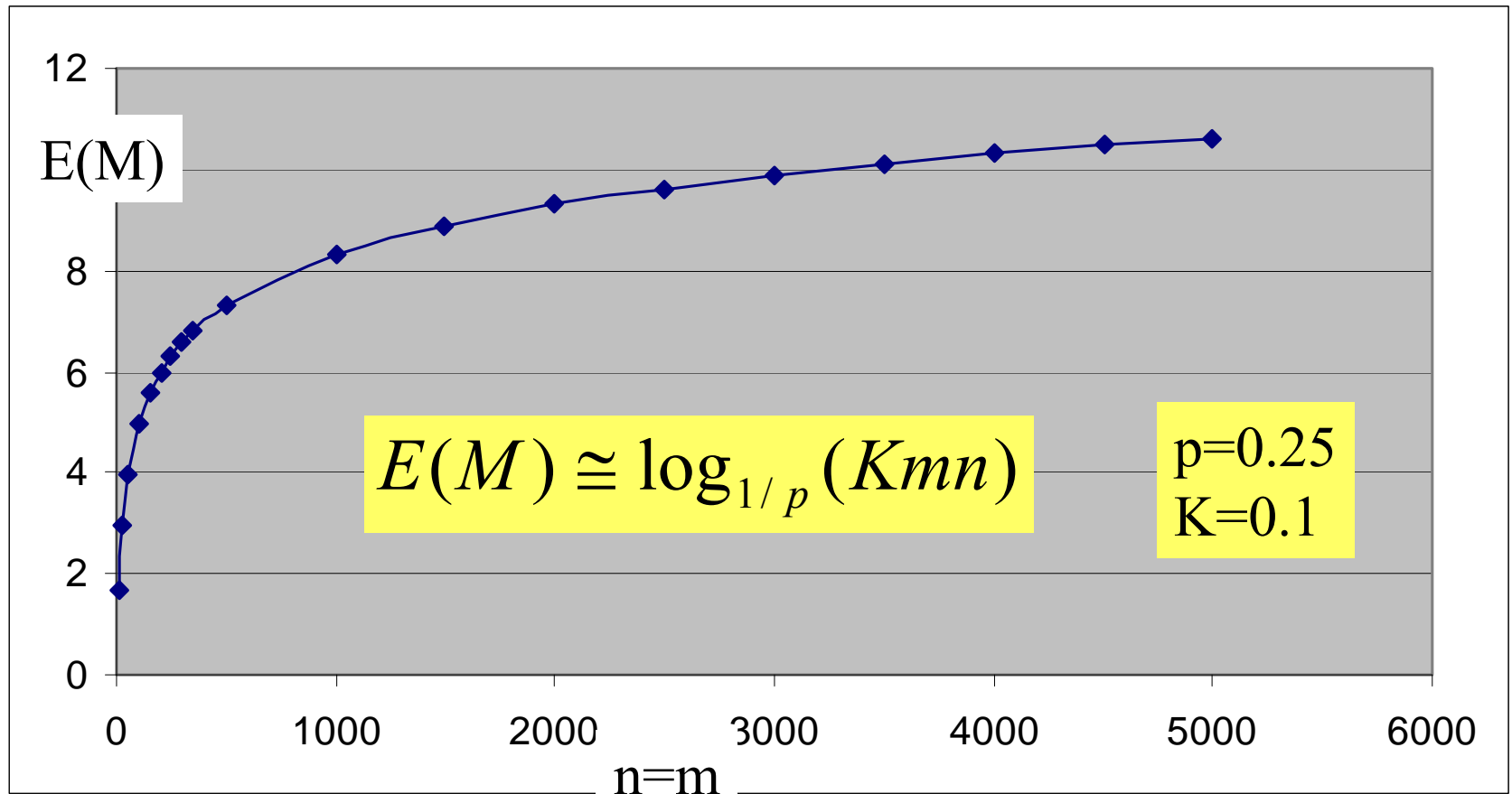
This equation can be simplified to

$$E(M) \cong \log_{1/p}(Kmn)$$

K is constant that depends p but not n and m .

(K accounts for the fact that there are not really $n \times m$ independent places to start alignment)

Mean of longest run



The mean of the highest possible alignment score is proportional to log of product of sequence lengths

Mean of longest run and λ

$$E(M) \cong \log_{1/p}(Kmn)$$

This equation can be rewritten by using $\lambda = \ln(1/p)$.

$$E(M) = \frac{\ln(Kmn)}{\lambda}$$

Instead of p , the scaling parameter λ is more commonly used in with scoring matrices.

K is of less influence (importance) than λ

Mean expected score $E(S)$

$$E(M) = \frac{\ln(Kmn)}{\lambda}$$

For alignments **scores** are preferred compared to maximum matching lengths.

IF $E(M)$ is known (region of longest match)

AND **scoring system** for matches/mismatches is given (e.g. PAM)

THEN **mean expected score** for a random sequence is

$$E(S) = \frac{\ln(Kmn)}{\lambda}$$

Note that for each scoring matrix the average score per match can be calculated

Mean expected score

$$E(S) = \frac{\ln(Kmn)}{\lambda}$$

Random sequence of coin tosses:

... TTT **HHHHHHH** TTT **HHH** TTT **H** T **HHHH** T **HH** T ...



Random alignment

... CAT **GGATGAC** CGTAAA ...
... ATA **GGATGAC** AAAAAA ...

Expected score for chosen scoring matrix = $E(S)$

Probability $P(S \geq x)$ and expected score $E(S \geq x)$

Question:

What is the probability $P(S \geq x)$?

Equivalent question:

Probability $P(S \geq x) = P_n(E(S \geq x))$

probability of observing score $S > x$ a number of times ($n \geq 1$)

Answer:

Use Poisson distribution

$$P(\mu = E(S \geq x)) = \frac{e^{-\mu} \mu^n}{n!}$$

Probability $P(S \geq x)$ and expected score $E(S \geq x)$

$$P(S \geq x) = P_n(\mu = E(S \geq x)) = \frac{e^{-\mu} \mu^n}{n!}$$

Probability of observing **one or more** segments with $S \geq x$:

$$P_{n \geq 1}(\mu = E(S \geq x)) = \sum_{n=1}^{\infty} \frac{e^{-\mu} \mu^n}{n!}$$

Probability $P(S \geq x)$ and expected score $E(S \geq x)$

Easier to calculate is:

$P(S < x) = P_{n=0}(E(S \geq x)) \rightarrow$ Probability of observing
no segments with $S \geq x$

$$P_{n=0}(\mu = E(S \geq x)) = e^{-\mu}$$

Note that $P(S \geq x) = 1 - P(S < x) \rightarrow P(S \geq x) = 1 - e^{-\mu}$

How do we calculate $\mu = E(S \geq x)$

Recall: longest run of heads

$$E(L) = n * p^L \Rightarrow E(R_L) = n * p^{R_L} = 1$$

run expected one

$$R = \log_{1/p}(n)$$

length of this run

$$R = \lceil \log_{1/p}(nm) \rceil$$

run for alignments

$$E(M) \cong \log_{1/p}(mn) + \log_{1/p}(1-p) + \gamma \log(e) - \frac{1}{2}$$

Mean run

$$E(M) \cong \log_{1/p}(Kmn) \quad \lambda = \ln(1/p)$$

Simplified

$$E(M) = \frac{\ln(Kmn)}{\lambda}$$

parameter λ

$$E(S) = \frac{\ln(Kmn)}{\lambda}$$

Mean score of longest matches
that are expected to occur once

Summarize: longest run of heads

$$E(L) = n * p^L \Rightarrow E(R_L) = n * p^{R_L} = 1$$

run expected one

$$R = \log_{1/p}(n)$$

length of this run

$$R = \lceil \log_{1/p}(nm) \rceil$$

run for alignments

$$E(M) \cong \log_{1/p}(mn) + \log_{1/p}(1-p) + \gamma \log(e) - \frac{1}{2}$$

Mean run

$$E(M) \cong \log_{1/p}(Kmn) \quad \lambda = \ln(1/p)$$

Simplified

$$E(M) = \frac{\ln(Kmn)}{\lambda}$$

parameter λ

$$E(S) = \frac{\ln(Kmn)}{\lambda}$$

Mean score of longest matches that are expected to occur once

Summarize: longest run of heads

$$E(L) = n * p^L \Rightarrow E(R_L) = n * p^{R_L} = 1$$

run expected one

$$E(S) = \frac{\ln(Kmn)}{\lambda}$$

$$E(S \geq x) = Kmn p^x$$

Score of at least x
is expected E times

Summarize: longest run of heads

$$Kmp^x = E$$

Score of at least $S=x$ is expected E times

$$\ln(Kmp^x) = \ln(E)$$

$$\ln(Kmn) + x \ln(p) = \ln(E)$$

$$\ln(Kmn) - x \ln(1/p) = \ln(E)$$

$$\ln(Kmn) - \lambda x = \ln(E)$$

$$E = e^{\ln(Kmn) - \lambda x}$$

$$E(S \geq x) = Kmne^{-\lambda x}$$

The expected number of random alignments with score $S \geq x$.

Probability $P(S \geq x)$ and expected score $E(S \geq x)$

Probability of getting any alignment whose score S is at least x

$$P(S \geq x) = 1 - e^{-\mu}$$

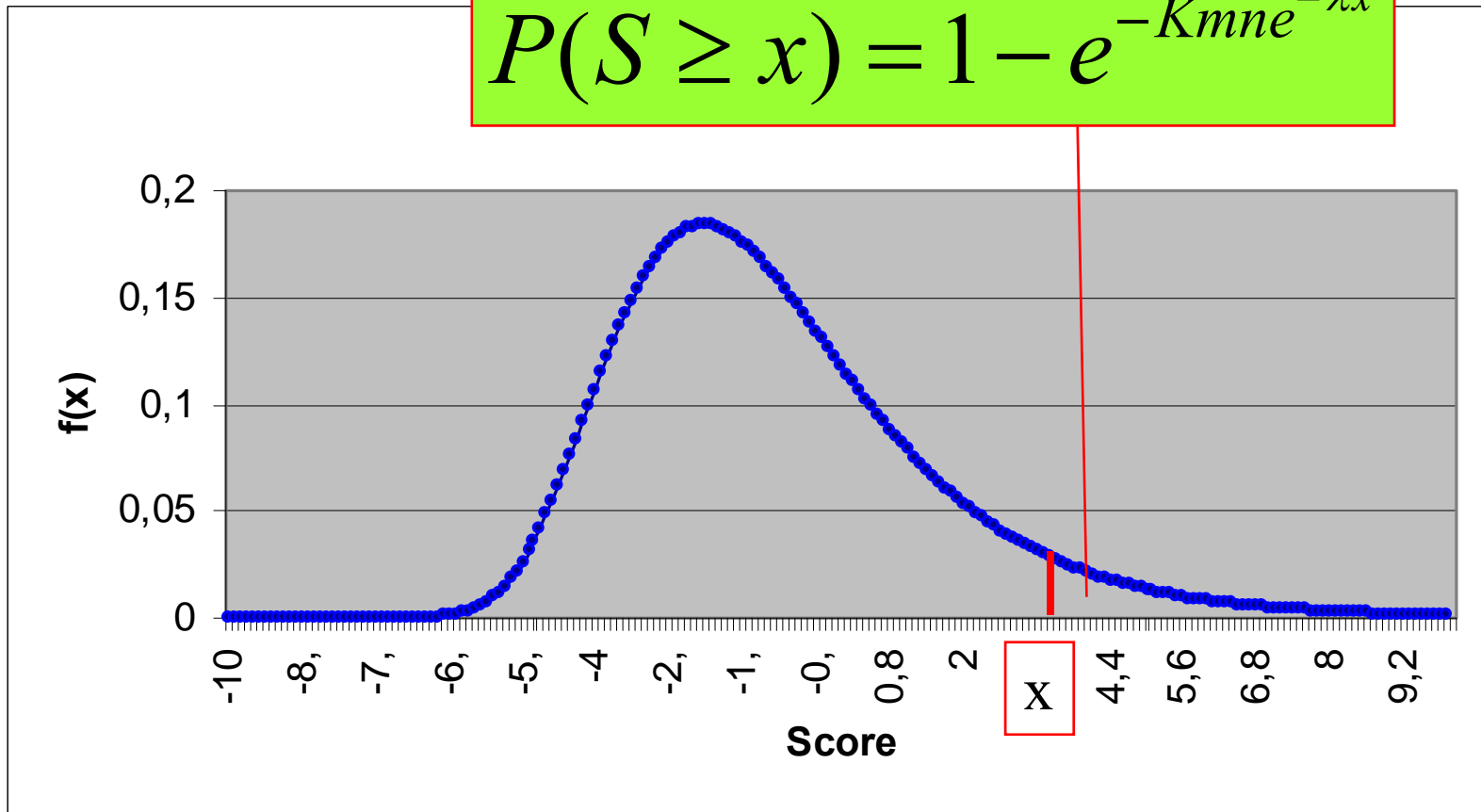
$$\mu = E(S \geq x) = Kmne^{-\lambda x}$$



$$P(S \geq x) = 1 - e^{-Kmne^{-\lambda x}}$$

Probability of observing score $S \geq x$ in random sequence

$$P(S \geq x) = 1 - e^{-Kmne^{-\lambda x}}$$



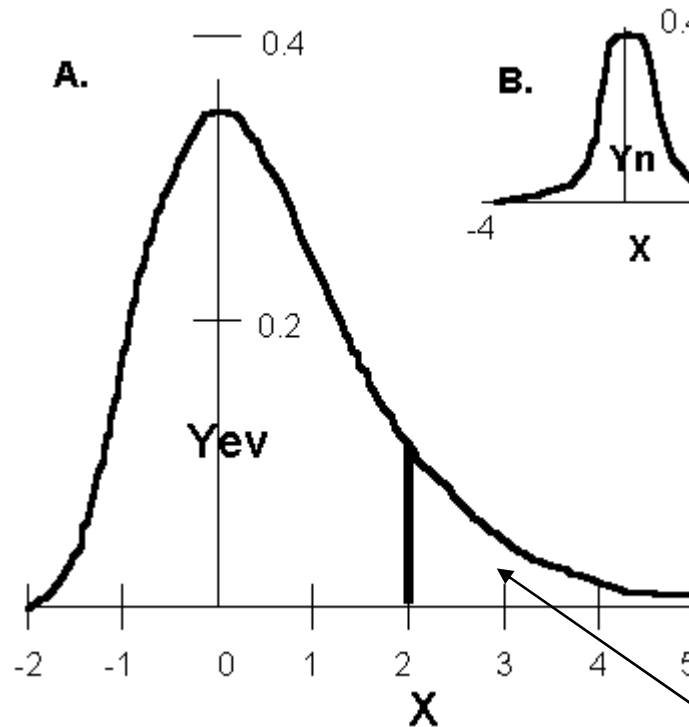
Extreme Value Distribution (EVD)

EVD and normal distribution

Standardized form

EVD

mean= $x=0.577$
variance= 1.64
area $-2 < x < +2 = 0.87$
area $-3 < x < +3 = 0.95$

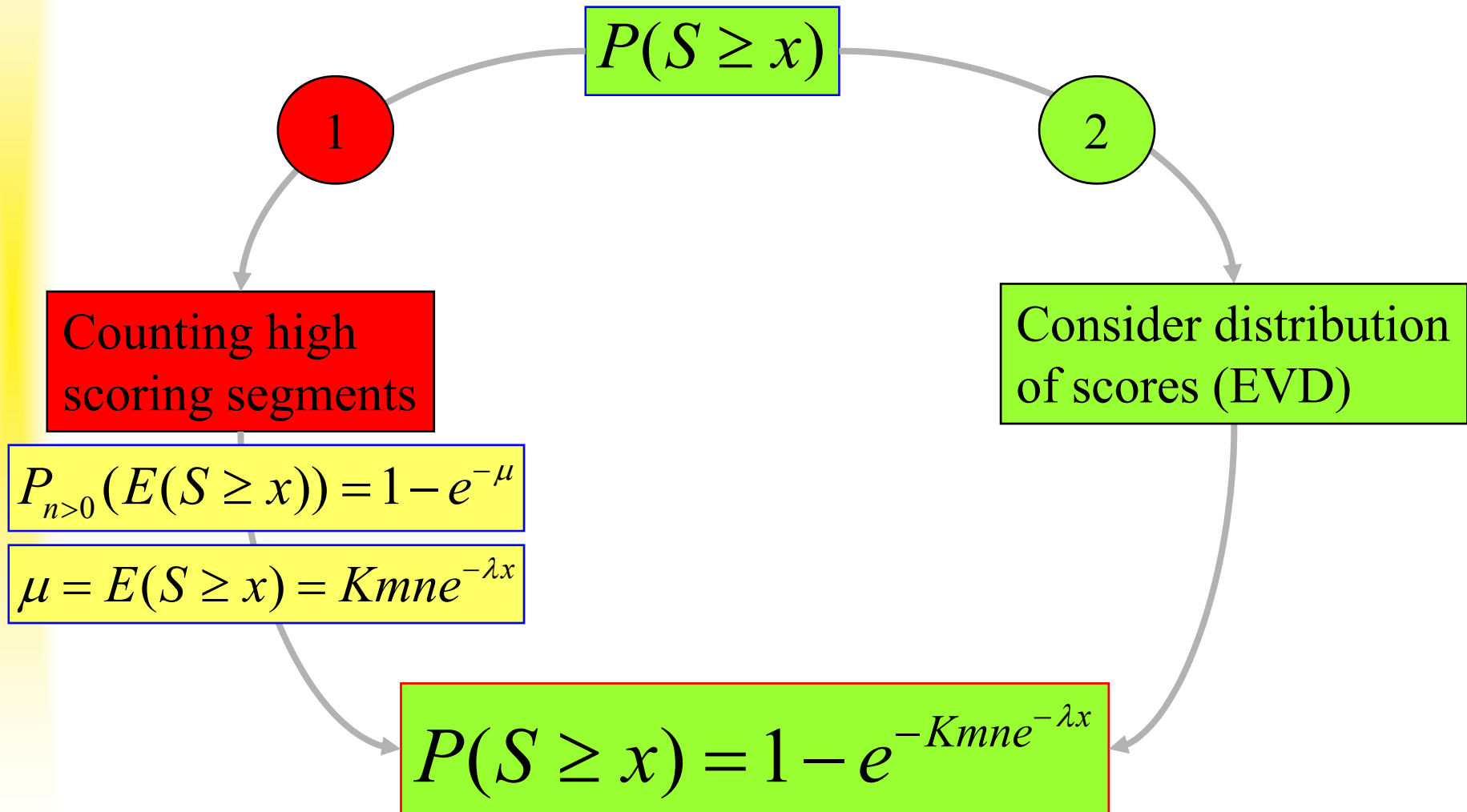


Normal distribution

mean= $x=0$
variance= 1
area $-2 < x < +2 = 0.954$

The scores must be greater than expected from the normal distribution to become statistically significant!

Calculating statistical significance of scores



Raw scores

Normalized scores

Bit scores

Expectation value

p-value

Raw scores

S_{raw}

Normalized scores

Bit scores

$$E(S \geq x) = Kmne^{-\lambda x}$$

Expectation value

$$P(S \geq x) = 1 - e^{-Kmne^{-\lambda x}}$$

p-value

Raw scores S_{raw}

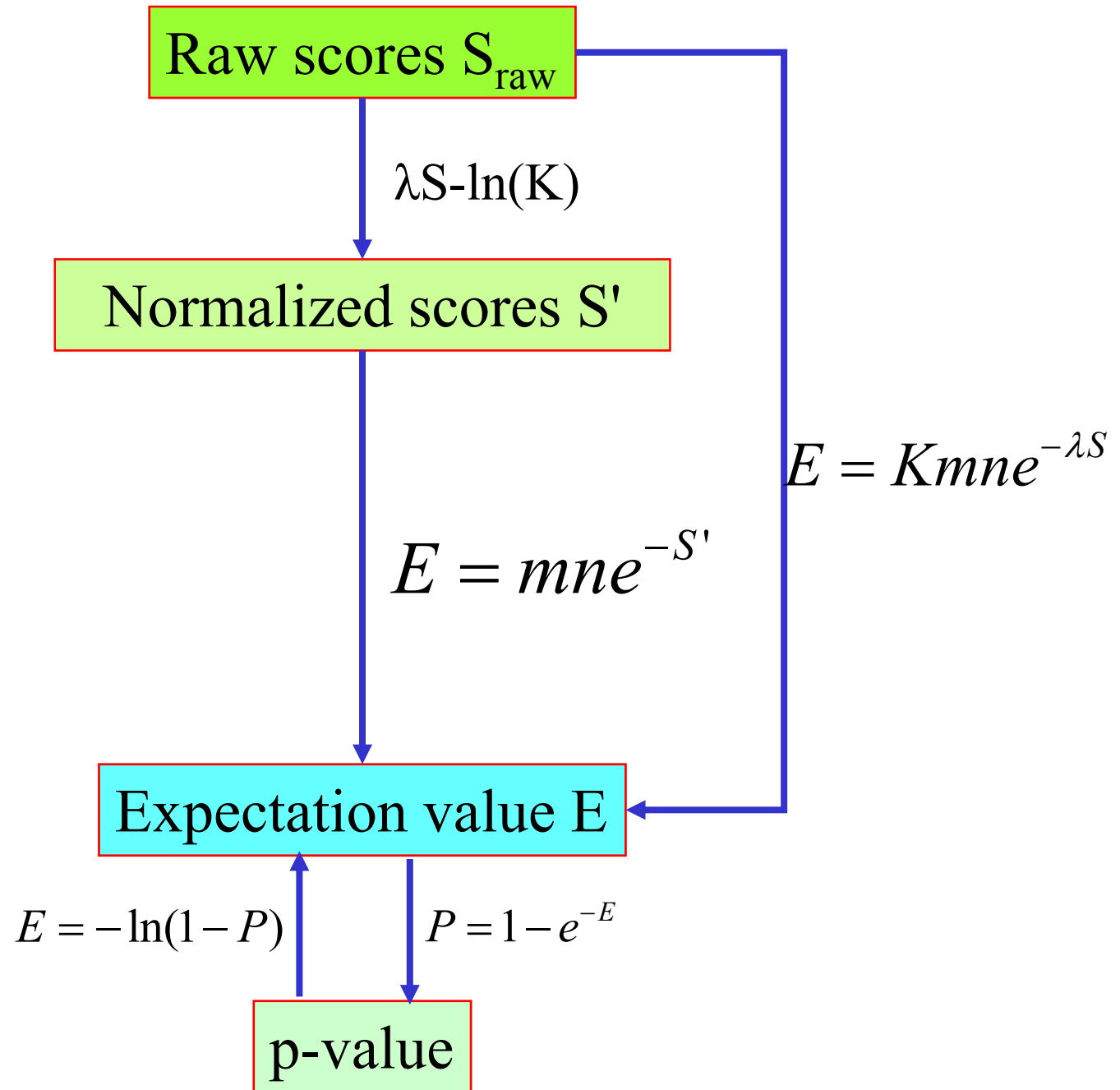
Expectation value E

p-value

$$E = Kmne^{-\lambda S}$$

$$E = -\ln(1 - P)$$

$$P = 1 - e^{-E}$$



Normalization of scores allows
comparison of scores

$$S_{\text{raw},1} = 100$$

$$S_{\text{raw},2} = 80$$

Does first score correspond to better alignment?

Answer: we can't say, unless we consider λ and K

Raw scores S_{raw}

$$\lambda S - \ln(K)$$

Normalized scores S'

$$S'_1 = 100$$

$$S'_2 = 80$$

Does first score correspond to better alignment?

Answer: yes.

Normalized scores S'

$$E(S_{raw} \geq x) = -Kmne^{-\lambda x}$$

$$P(S_{raw} \geq x) = 1 - e^{-Kmne^{-\lambda x}}$$

$$S' = \lambda S_{raw} - \ln(K)$$

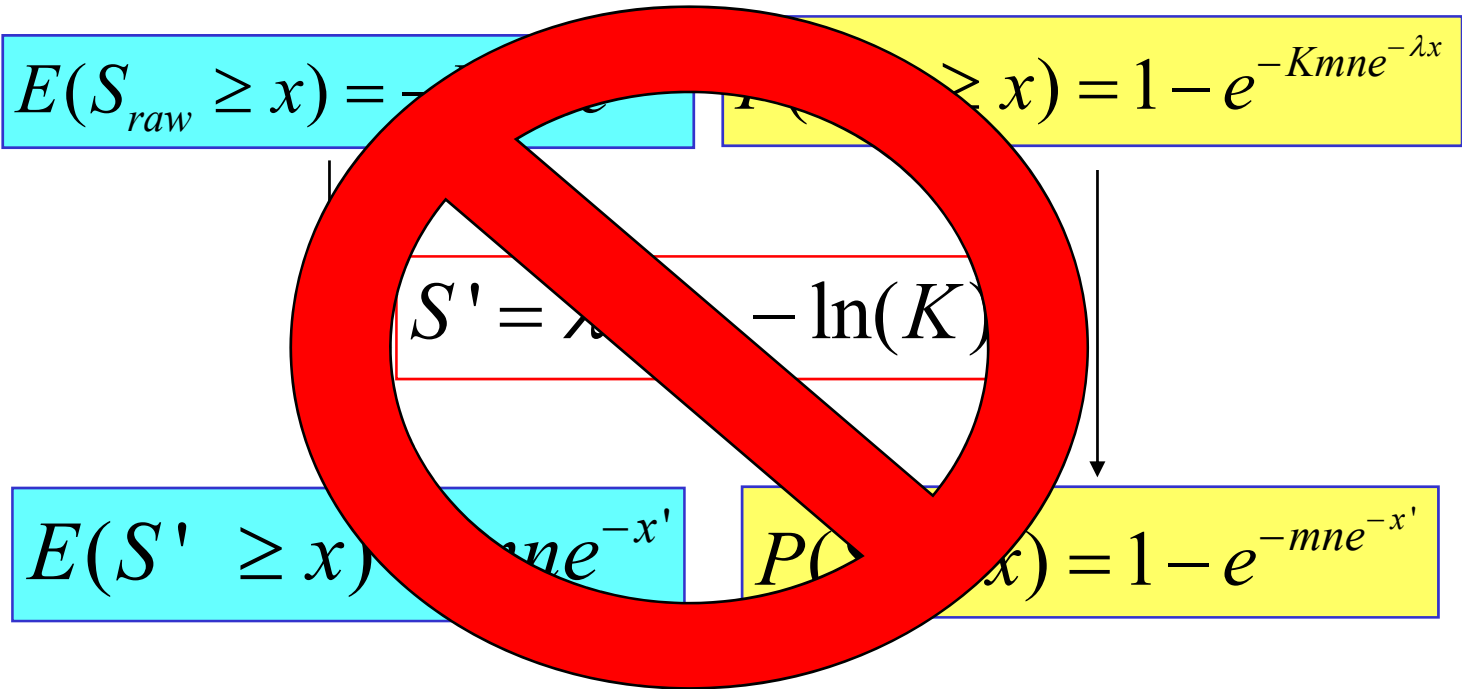
$$E(S' \geq x) = mne^{-x'}$$

$$P(S' > x) = 1 - e^{-mne^{-x'}}$$

note x'

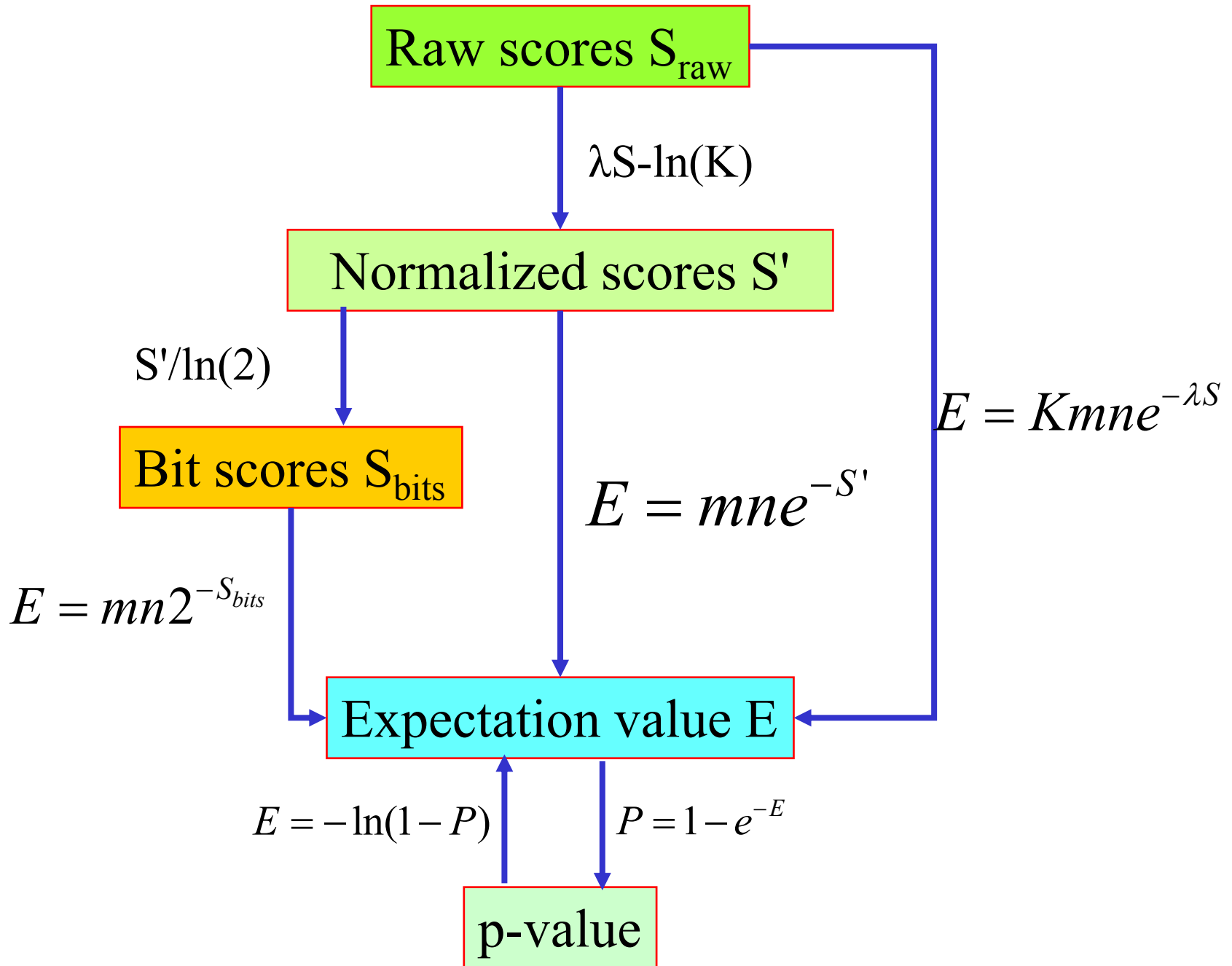
S' is independent of scoring scheme

Normalized scores S'



note x'

S' is independent of scoring scheme



Normalized bit-scores S'

$$E(S_{raw} \geq x) = -K m n e^{-\lambda x}$$

$$P(S_{raw} \geq x) = 1 - e^{-K m n e^{-\lambda x}}$$

$$S_{bit} = \frac{\lambda S_{raw} - \ln(K)}{\ln 2}$$

$$E(S_{bits} \geq x) = m n 2^{-x'}$$

$$P(S' > x) = 1 - e^{-m n 2^{-x'}}$$

normalized score expressed as bits

Bits are the common 'currency' for scores

Raw scores S_{raw}

$$\lambda S - \ln(K)$$

Normalized scores S'

$$S' / \ln(2)$$

Bit scores S_{bits}

$$E = mn2^{-S_{\text{bits}}}$$

$$E = mne^{-S'}$$

$$E = Kmne^{-\lambda S}$$

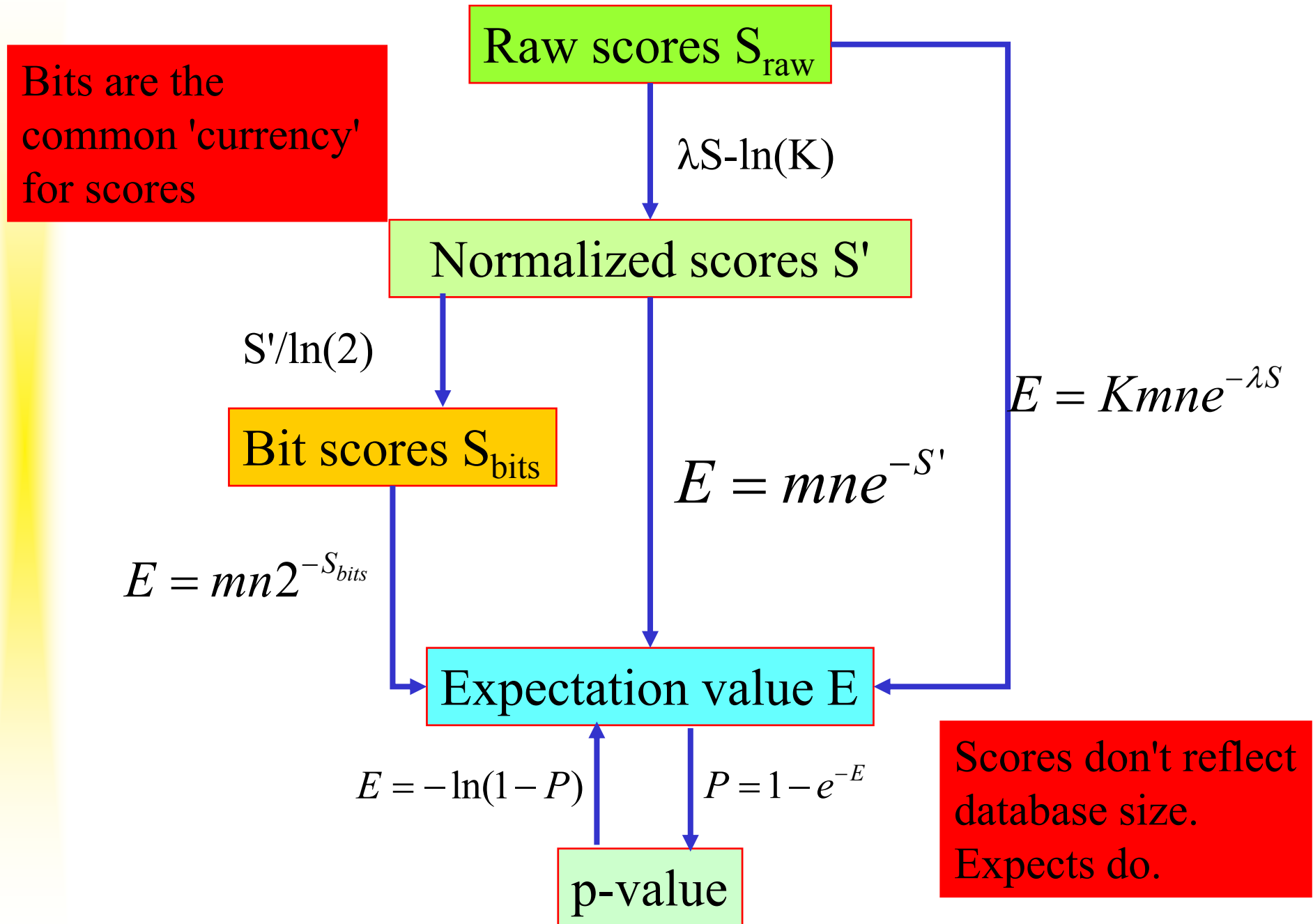
Expectation value E

$$E = -\ln(1 - P)$$

$$P = 1 - e^{-E}$$

p-value

Scores don't reflect database size. Expects do.



PAM and BLOSUM matrix
contain bit-scores

Bit scores S_{bits}

$$E = mn2^{-S_{bits}}$$

Expectation value E

$$P = 1 - e^{-E}$$

p-value

The scaling factor λ

PAM and BLOSUM scores were derived from target specific models of evolution as

$$s_{ij} = \log \left(\frac{q_{ij}}{p_i p_j} \right)$$

p: background frequencies

q: target frequencies in alignments

For PAM and BLOSUM p and q are obtained from observations and therefore **represent real probabilities**

The scaling factor λ

$$s_{ij} = \log\left(\frac{q_{ij}}{p_i p_j}\right) \longrightarrow q_{ij} = p_i p_j e^{s_{ij}}$$

For arbitrary scoring matrices, q may not represent true probabilities. Therefore, we have to scale the scores with λ

$$q_{ij} = p_i p_j e^{\lambda s_{ij}} \longrightarrow \text{Sum}(q)=1$$

The scaling factor λ

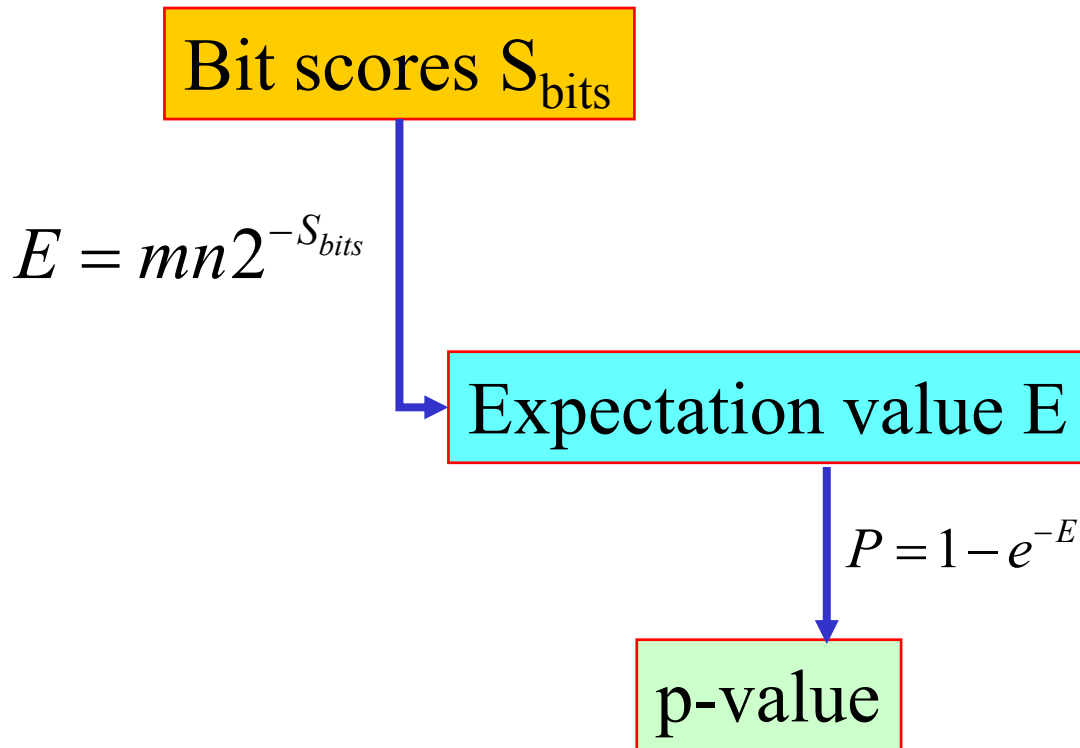
$$q_{ij} = p_i p_j e^{\lambda s_{ij}} \longrightarrow s_{ij} = \frac{\log\left(\frac{q_{ij}}{p_i p_j}\right)}{\lambda}$$

We can also 'rescale' the PAM/BLOSUM matrices to unit of bits $\rightarrow \lambda = \ln(2)$

$$s_{ij} = \frac{\log\left(\frac{q_{ij}}{p_i p_j}\right)}{\lambda} = \frac{\log\left(\frac{q_{ij}}{p_i p_j}\right)}{\ln(2)} = \log_2\left(\frac{q_{ij}}{p_i p_j}\right)$$

$$q_{ij} \neq p_i p_j e^{s_{ij}}$$
$$q_{ij} = p_i p_j e^{\ln(2) s_{ij}}$$

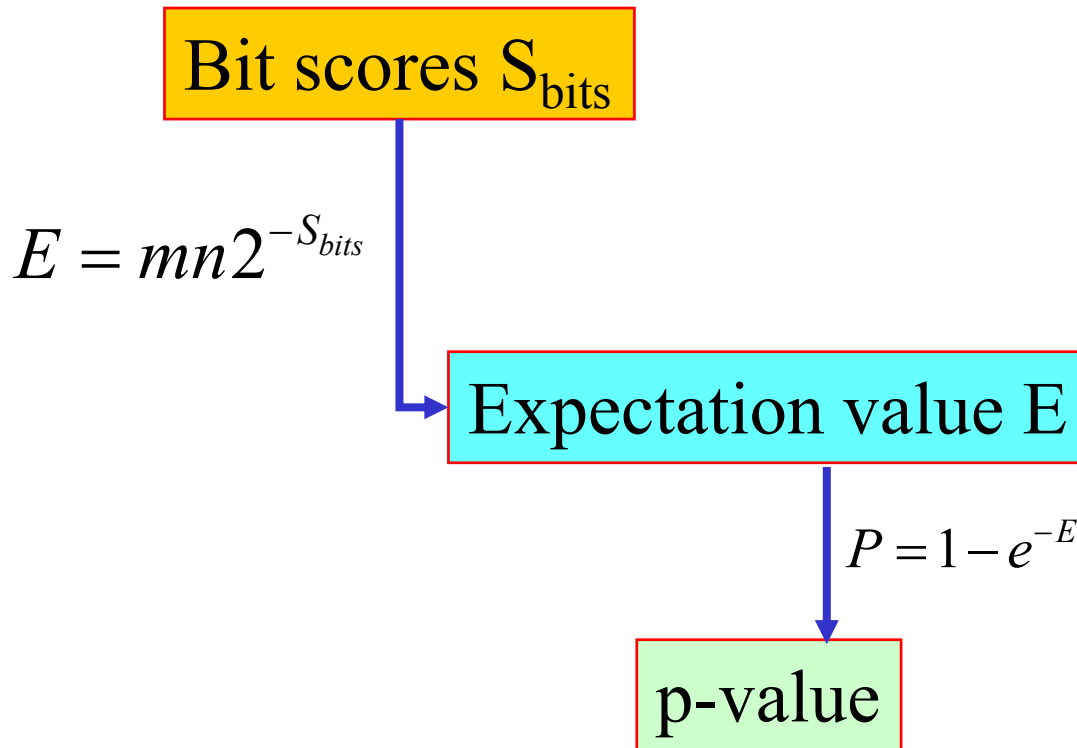
PAM and BLOSUM matrix are already normalized to bit-scores



Note that:

IF S_{bits} gets larger
THEN E gets smaller
THEN e^{-E} gets larger
THUS $p=1 - e^{-E}$ gets smaller (more significant)

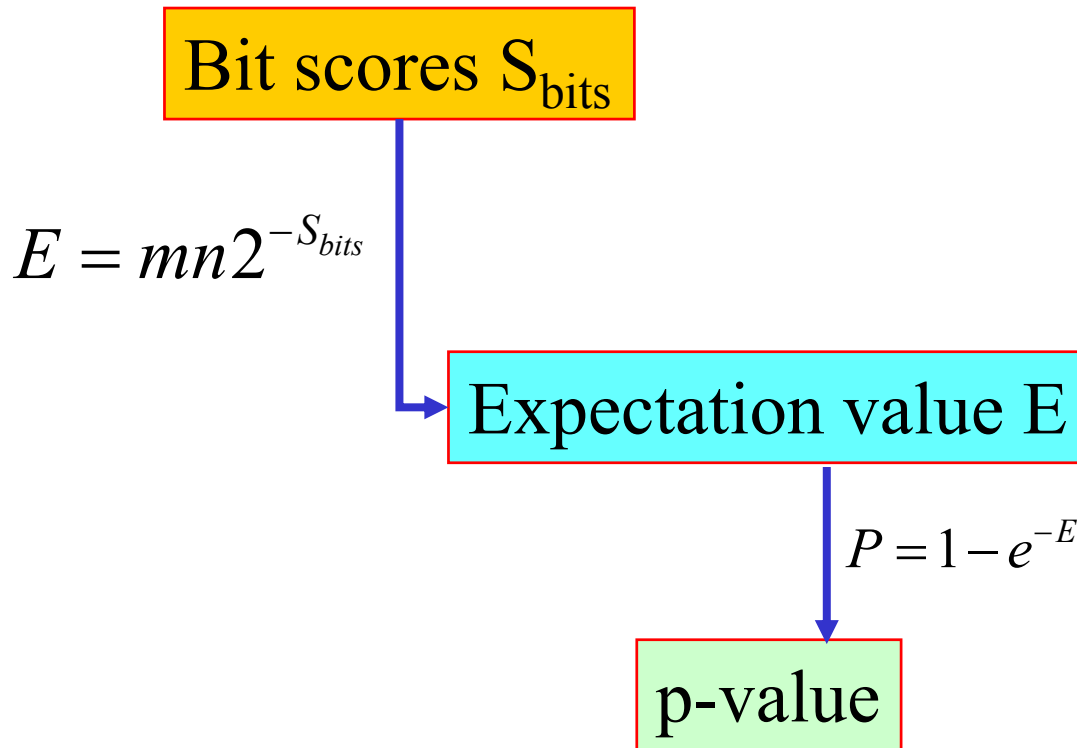
We already knew this!



Note also that:

IF We have a score S_{bits}
AND if the database ($n*m$) grows
THEN E gets larger
THEN e^{-E} gets smaller
THUS $p=1 - e^{-E}$ gets larger (less significant)

Thus a significant hit today may not be significant tomorrow!



Quick determination of alignment score

For typical scoring matrices and protein sequences $K=0.1$

$$P(S \geq x) \sim Kmn e^{-\lambda x}$$

For $S \gg x$ $P \sim E$

The probability that a random alignment reaches score S :

$${}^2\log(P) = {}^2\log(Kmn e^{-\lambda S}) = {}^2\log(Kmn) - S$$

Rearrange equation and choosing $K=0.1$ and **$P=0.05$**

$$S = {}^2\log(Kmn) - {}^2\log P$$

$$S \sim {}^2\log(nm)$$

Example

Suppose 2 protein sequences of 250 AA are aligned.

The following alignment is found

```
FWLEVEGNSMTAPTG
FWLDVQGDSMTAPAG
```

A significant alignment of 2 random sequences will at least have a score of $S \sim {}^2\log(nm) = {}^2\log(250*250) = 16$ bits

PAM250 score = 75 (1/3 bits) = 25 bits

$25 > 16 \rightarrow$ alignment is significant at $P=0.05$

Example continued

Score from PAM250 is $S=75$

For PAM250 $K=0.09$ and $\lambda=0.229$

Normalized score

$$S' = 0.229*75 - \ln(0.09*250*250)$$
$$S' = 8.57$$

$$P(S' > 8.57) = 1 - \exp[-e^{-8.57}] = 1.9 * 10^{-4}$$

This is the probability of obtaining this scoring in a random alignment of 2 sequences of length=250

Scoring matrices
from an
information theoretic perspective

Information theory

- **Information theory:** deals with messages that are numerically coded for their description, storage and communication.
- In our case: message is DNA or protein sequence

Definition:

$$I = -\log_2(P)$$

Unit: bits

I == Information

P == Probability

Relative entropy

- The score expressed as ‘bits’

$$s_{ij} = -2 \log\left(\frac{q_{ij}}{p_i p_j}\right)$$

- The average score (information) per residue pair in an alignment is called the **relative entropy**

$$H = \sum_{i,j} q_{ij} s_{ij} = \sum_{i,j} q_{ij} -2 \log\left(\frac{q_{ij}}{p_i p_j}\right)$$

Comparing sequences

Comparing protein of 250 AA

against

SwissProt database 40.000.000 AA

$$S \sim 2 \log(nm)$$

we need $\text{LOG}_2(250 * 40.000.000) = 33$ bits of information

Relative entropy for substitution matrices

- For a PAM matrix we can calculate the relative entropy:

$$H = \sum_{i,j} q_{ij} s_{ij} = \sum_{i,j} q_{ij}^2 \log \left(\frac{q_{ij}}{p_i p_j} \right) \quad \text{information per residue}$$

$$S \sim 2 \log(nm) \quad \text{information required}$$

—————→ Minimal length of segment that can be distinguished from chance:

$$\text{Length} = S/H$$

Relative entropy for PAM matrices

| PAM distance | H(bits) | Min. significant length (33 bits) | PAM distance | H(bits) | Min. significant length (33 bits) |
|--------------|---------|-----------------------------------|--------------|---------|-----------------------------------|
| 0 | 4,17 | 8 | 180 | 0,6 | 55 |
| 10 | 3,43 | 10 | 190 | 0,55 | 60 |
| 20 | 2,95 | 11 | 200 | 0,51 | 65 |
| 30 | 2,57 | 13 | 210 | 0,48 | 69 |
| 40 | 2,26 | 15 | 220 | 0,45 | 73 |
| 50 | 2,00 | 17 | 230 | 0,42 | 79 |
| 60 | 1,79 | 18 | 240 | 0,39 | 85 |
| 70 | 1,60 | 21 | 250 | 0,36 | 92 |
| 80 | 1,44 | 23 | 260 | 0,34 | 97 |
| 90 | 1,30 | 25 | 270 | 0,32 | 103 |
| 100 | 1,18 | 28 | 280 | 0,3 | 110 |
| 110 | 1,08 | 31 | 290 | 0,28 | 118 |
| 120 | 0,98 | 34 | 300 | 0,27 | 113 |
| 130 | 0,9 | 37 | 310 | 0,25 | 120 |
| 140 | 0,82 | 40 | 320 | 0,24 | 127 |
| 150 | 0,76 | 43 | 330 | 0,22 | 134 |
| 160 | 0,7 | 47 | 340 | 0,21 | 141 |
| 170 | 0,65 | 51 | 350 | 0,2 | 149 |

Example

- For finding a significant segment alignment of a protein of 250 residues against the swissprot database we need at least 33 bits of information
- PAM 10 3,43 bits/AA 10 residues required
- PAM120 0,98 bits/AA 34 residues required
- PAM250 0,36 bits/AA 92 residues required
- Thus, if the divergence between proteins becomes larger one needs a larger segment

Putting it all together:

BLAST

BLAST output

output>gi | 6006425 | emb | CAB56829.1 | hemoglobin alpha chain

Length = 142

Score = 33.9 bits (76), Expect = 0.66

Identities = 15/15 (100%), Positives = 15/15 (100%)

```
Query: 1  MVLSAADKGNVKA AW 15
      MVLSAADKGNVKA AW
Sbjct: 1  MVLSAADKGNVKA AW 15
```

S_{raw}

Database: All non-redundant GenBank CDS

Number of letters in database: 436,700,696

Number of sequences in database: 1,364,053

Gapped

Lambda K H

0.267

0.0410

0.140

BLAST output

output>gi|6006425|emb|CAB56829.1| hemoglobin alpha chain

Length = 142

Score = 33.9 bits (76), Expect = 0.66

Identities = 15/15 (100%), Positives = 15/15 (100%)

```
Query: 1  MVLSAADKGNVKA AW 15
      MVLSAADKGNVKA AW
Sbjct: 1  MVLSAADKGNVKA AW 15
```

S_{raw}

Database: All non-redundant GenBank CDS

Number of letters in database: 436,700,696

Number of sequences in database: 1,364,053

Gapped

Lambda K H

0.267

0.0410

0.140

$$S' = \frac{\lambda S - \ln K}{\ln 2}$$

BLAST output

output>gi|6006425|emb|CAB56829.1| hemoglobin alpha chain

Length = 142

Score = 33.9 bits (76), Expect = 0.66

Identities = 15/15 (100%), Positives = 15/15 (100%)

```
Query: 1  MVLSAADKGNVKA AW 15
      MVLSAADKGNVKA AW
Sbjct: 1  MVLSAADKGNVKA AW 15
```

$$E = mn2^{-S'}$$

S_{raw}

Database: All non-redundant GenBank CDS

Number of letters in database: 436,700,696

Number of sequences in database: 1,364,053

Gapped

Lambda K H

0.267

0.0410

0.140

$$S' = \frac{\lambda S - \ln K}{\ln 2}$$

BLAST output

output>gi | 6006425 | emb | CAB56829.1 | hemoglobin alpha chain

Length = 142

Score = 33.9 bits (76), Expect = 0.66

Identities = 15/15 (100%), Positives = 15/15 (100%)

```
Query: 1  MVLSAADKGNVKA AW 15
        MVLSAADKGNVKA AW
Sbjct: 1  MVLSAADKGNVKA AW 15
```

$$E = mn2^{-S'}$$

S_{raw}

$$S' = \frac{\lambda S - \ln K}{\ln 2}$$

Database: All non-redundant GenBank CDS

Number of letters in database: 436,700,696

Number of sequences in database: 1,364,053

Gapped

Lambda K H

0.267

0.0410

0.140

$$H = \sum q_{ij}^2 \log \left(\frac{q_{ij}}{p_i p_j} \right)$$

BLAST output

output>gi|6006425|emb|CAB56829.1| hemoglobin alpha chain

Length = 142

Score = 33.9 bits (76), Expect = 0.66

Identities = 15/15 (100%), Positives = 15/15 (100%)

```
Query: 1  MVLSAADKGNVKA AW 15
      MVLSAADKGNVKA AW
Sbjct: 1  MVLSAADKGNVKA AW 15
```

$$E = mn2^{-S'}$$

S_{raw}

Database: All non-redundant GenBank CDS

Number of letters in database: 436,700,696

Number of sequences in database: 1,364,053

Gapped

| Lambda | K | H |
|--------|--------|-------|
| 0.267 | 0.0410 | 0.140 |

$$S' = \frac{\lambda S - \ln K}{\ln 2}$$

$$H = \sum q_{ij}^2 \log \left(\frac{q_{ij}}{p_i p_j} \right)$$

BLAST output

output>gi | 6006425 | emb | CAB56829.1 | hemoglobin alpha chain

Length = 142

Score = 33.9 bits (76), Expect = 0.66

Identities = 15/15 (100%), Positives = 15/15 (100%)

```
Query: 1  MVLSAADKGNVKA AW 15
       MVLSAADKGNVKA AW
Sbjct: 1  MVLSAADKGNVKA AW 15
```

$$E = mn2^{-S'}$$

S_{raw}

Database: All non-redundant GenBank CDS

Number of letters in database: 436,700,696

Number of sequences in database: 1,364,053

Gapped

| Lambda | K | H |
|--------|--------|-------|
| 0.267 | 0.0410 | 0.140 |

$$S' = \frac{\lambda S - \ln K}{\ln 2}$$

$$H = \sum q_{ij}^2 \log \left(\frac{q_{ij}}{p_i p_j} \right)$$

Where is the p-value???

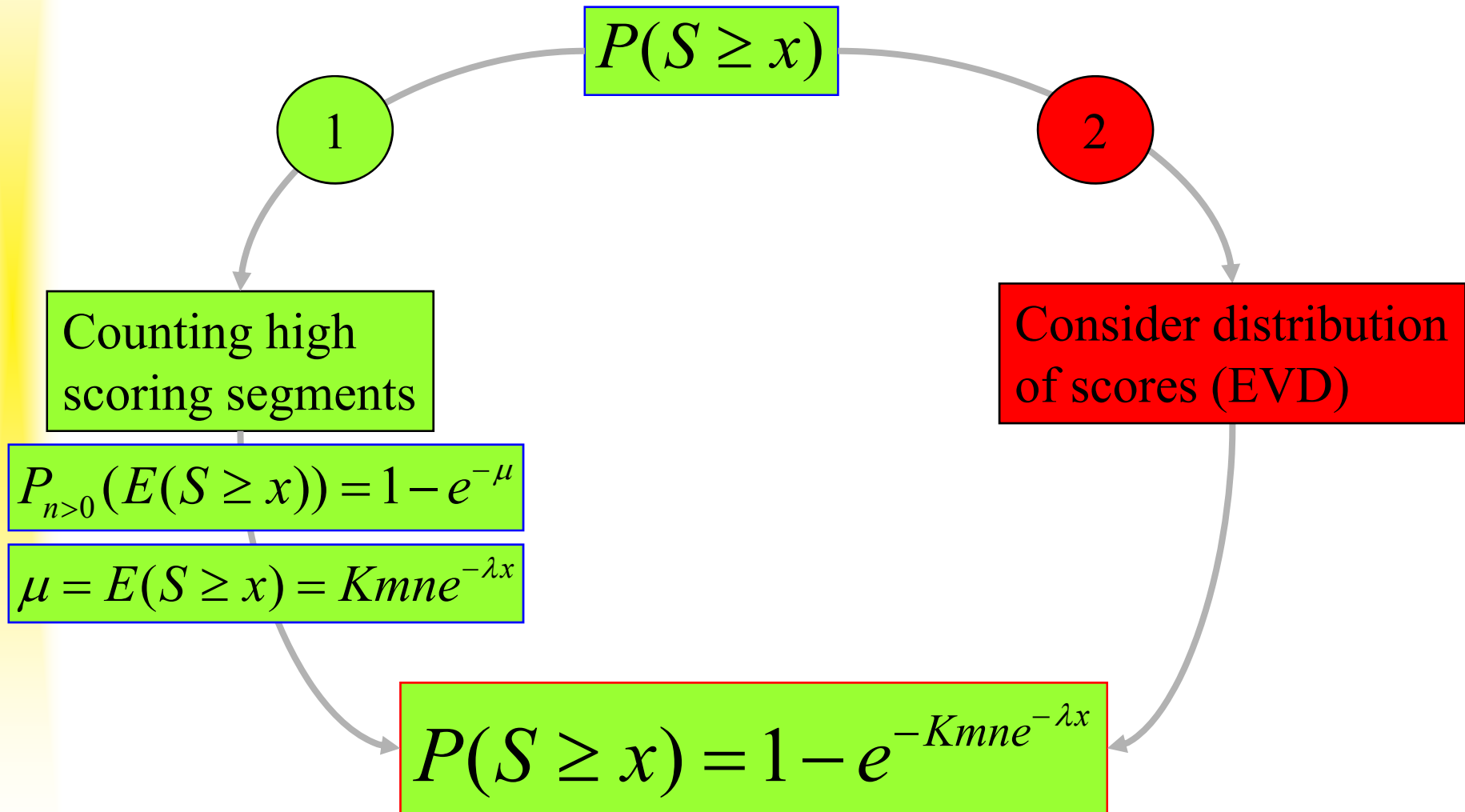
Example

- If one expects to find three segments with $S \geq x$, then $P(S \geq x) = 1 - e^{-3} = 0.95$
- The BLAST programs report E-values rather than P-values because it is easier to understand the difference between for example
 - E value of 5 and 10 than
 - P-values of 0.993 and 0.99995
- If $E < 0.01$ then P-values and E-values are nearly identical.



Are there any parts you didn't understand???

Calculating statistical significance of scores



Alignment scores and Extreme Value Distribution

Sequence alignments may result in both good and bad alignments (high and low scores).

Most biologically interesting alignments are those that give the highest score.

These 'highest scores' follow an [Extreme Value Distribution](#) (EVD)

Alignment "experiment"

Experiment:

- **Generate set of random alignments**
- **Keep highest score S**
- **Repeat this procedure N times**

Goal of this experiment:

determine probability that the score of a random alignment reaches the score of an alignment of two real sequences.

If this probability is very low than real alignment is significant

Maximum scores

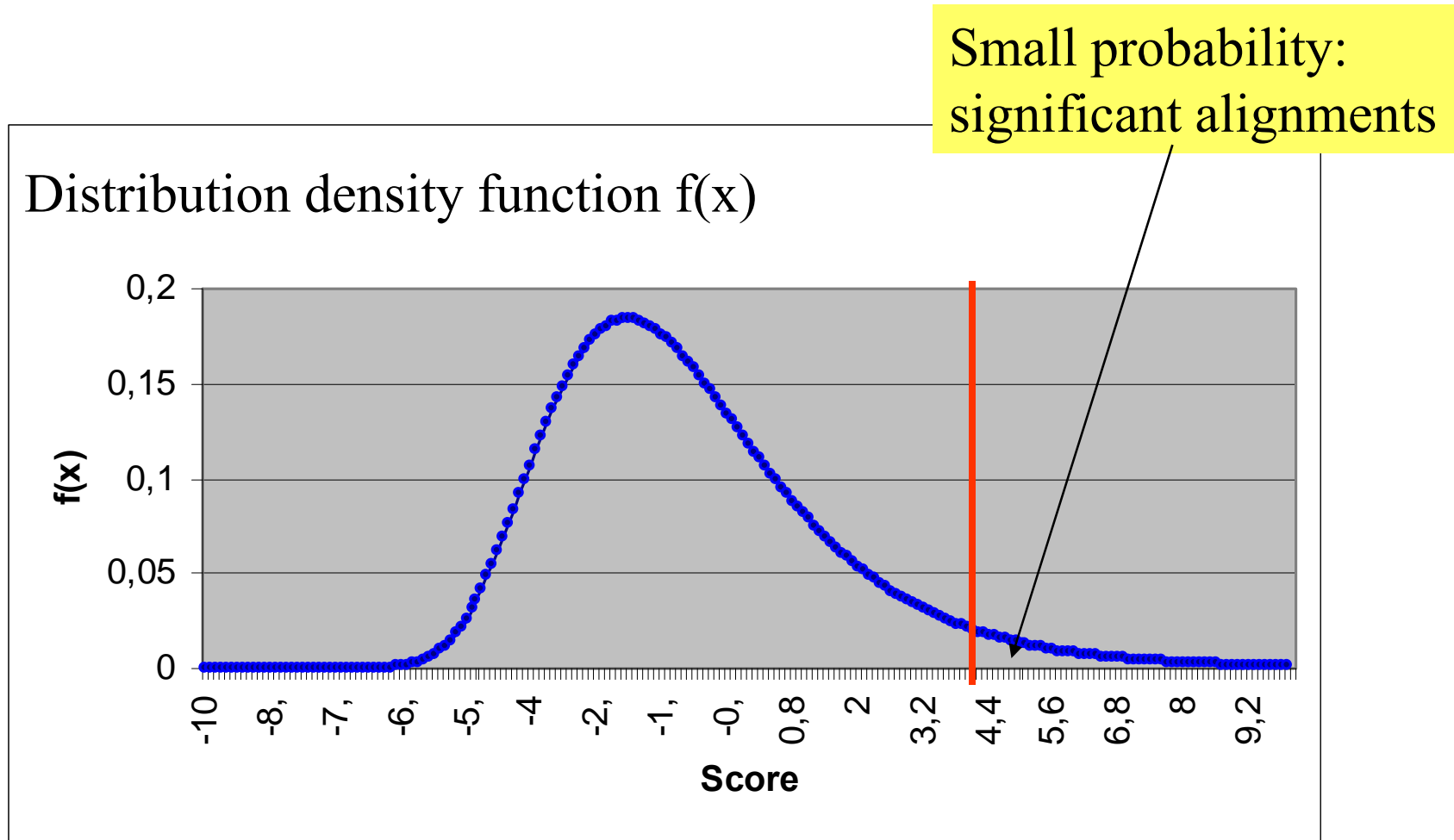
Experiment 1: 100 alignments ($p=0.25$, n,m):

| | |
|---------------------------------|------|
| Toss 1: H H T T H H.....H T | S=7 |
| Toss 2: T H H H H TT T | S=12 |
| | |
| | |
| Toss 100: T T H H H H H ... H H | S=8 |

Maximum = 12

The maximum scores follow extreme value distribution

EVD for random alignments



The Extreme Value Distribution

Probability density function

$$f(x') = \frac{1}{\beta} e^{-\frac{x'-\alpha}{\beta}} \exp\left[-e^{-\frac{x'-\alpha}{\beta}}\right]$$

α =mode
 β =scale parameter

Standardized form

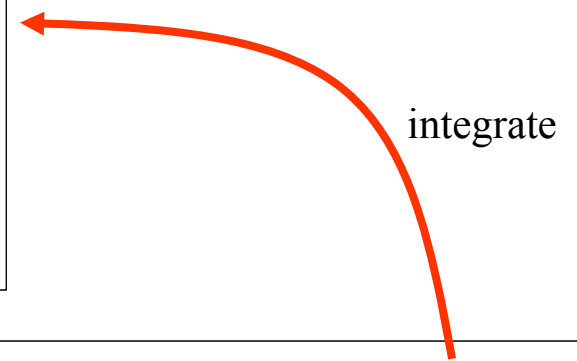
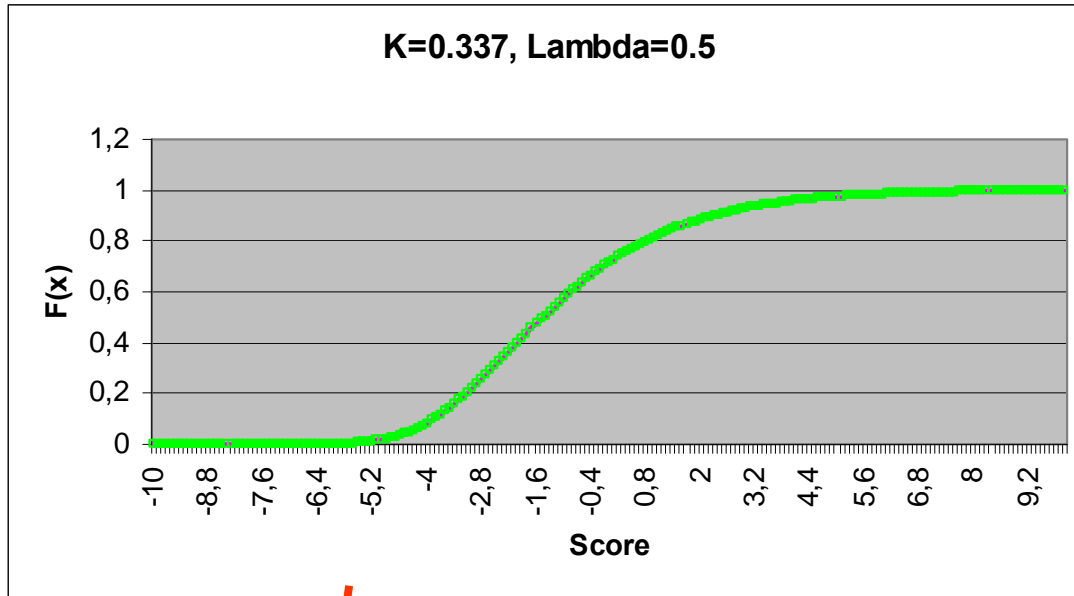
$$f(x) = e^{-x} \exp[-e^{-x}]$$

Cummulative distribution function

$$P(S < x) = \exp\left[-e^{-\frac{x'-\alpha}{\beta}}\right] \quad \text{or} \quad P(S < x) = \exp[-e^{-x}]$$

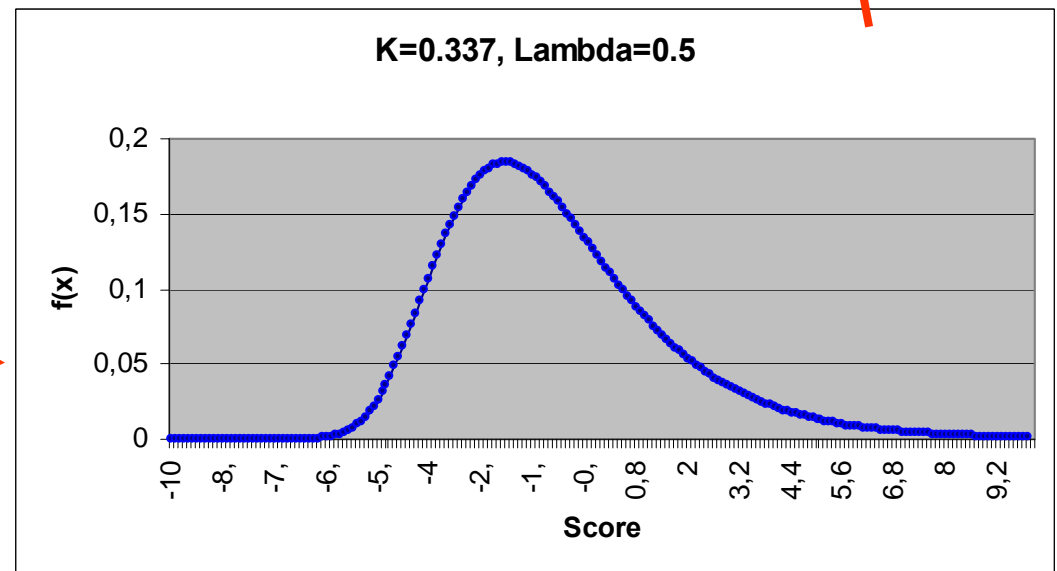
EVD

Cummulative distribution function



differentiate

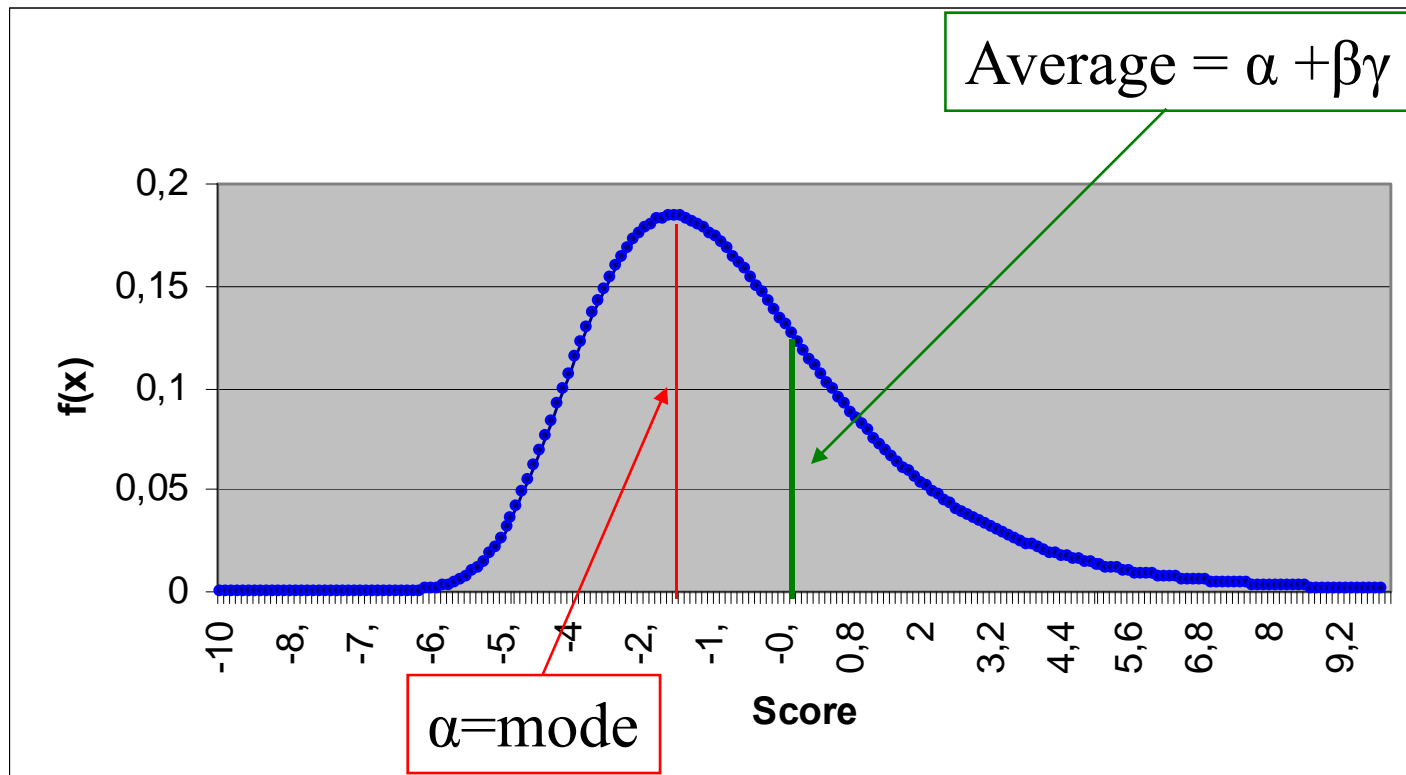
Probability function
(area = 1)



EVD

$$f(x') = \frac{1}{\beta} e^{-\frac{x'-\alpha}{\beta}} \exp\left[-e^{-\frac{x'-\alpha}{\beta}}\right]$$

$$std.dev = \sqrt{\frac{\beta^2 \pi^2}{6}}$$

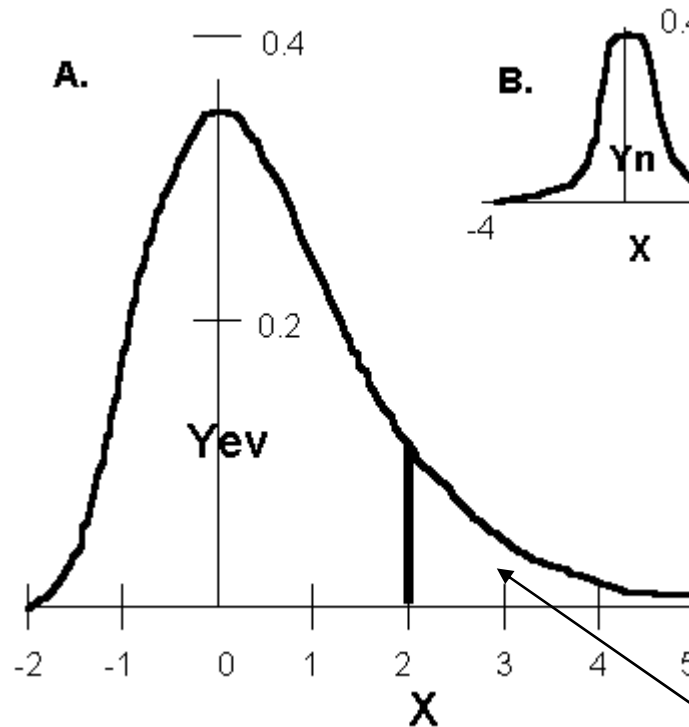


EVD and normal distribution

Standardized form

EVD

mean= $x=0.577$
variance= 1.64
area $-2 < x < +2 = 0.87$
area $-3 < x < +3 = 0.95$



Normal distribution

mean= $x=0$
variance= 1
area $-2 < x < +2 = 0.954$

The scores must be greater than expected from the normal distribution to become statistically significant!

P(S>=x) from the EVD

$$P(S < x) = \exp\left[-e^{-\frac{x-\alpha}{\beta}}\right]$$

Cummulative distribution

$$P(S \geq x) = 1 - \exp\left[-e^{-\lambda(x-\alpha)}\right]$$

$$\lambda = 1/\beta$$

$$\alpha = E(S) = \frac{\ln(Kmn)}{\lambda}$$

E(S) = mean score of longest matches that are expected to occur once

$$P(S \geq x) = 1 - e^{-Kmn e^{-\lambda x}}$$

(λ and α can easily be calculated from mean and std)

Calculating statistical significance of scores

