# Score and Significance

# Illustration

- Try BLAST the following protein:
- >pdb|6WPT|D Chain D, S309 neutralizing antibody heavy chain
  QVQLVQSGAEVKKPGASVKVSCKASGYPFTSYGISWVRQAPGQGLEWMGWISTYNGNTNYAQKFQGRVTM
  TTDTSTTTGYMELRRLRSDDTAVYYCARDYTRGAWFGESLIGGFDNWGQGTLVTVSS
- See the description of this sequence at https://www.ncbi.nlm.nih.gov/protein/6WPT_D

# Optimization Problem

- **Instance**: describes the input
- **Feasible Solution**: describes the format of the output
- **Score Function**: measures how good a solution is
- **Objective**: either maximize or minimize the score.

E.g. Sequence Alignment

- **Instance**: two sequences S and T
- **Feasible Solution**: insert gaps into S and T so that they have the same length
- **Score Function**: add up the score of each column
- **Objective**: to maximize the score

# Purposes of the Score

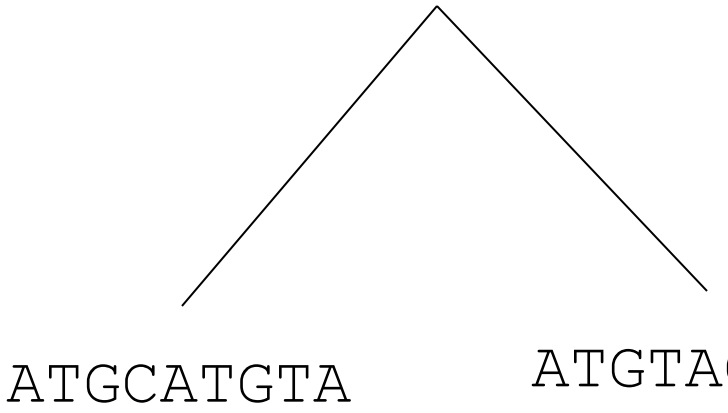
- Purpose 1: It helps us to compare solutions of the same instance.
  - Which alignment is the best for the same input (s,t).
- Purpose 2: It helps us to compare solutions of different instances.
  - Which of (s,t) and (x,y) is more likely to be a homology?
- Purpose 3: It helps us to tell how significant the solution is.
  - Does the alignment between s and t indicate that they are homologous?

Which solution is better? – same instance.

# PURPOSE 1

# Purpose 1

- We first examine how the scoring function is designed for the first purpose – compare two alignments and tell which one is better.
- Recall that what we really want is to find out homologies.




```
ATGCA-TGTA
|||  |  ||  |
ATGTACTG-A
```

# An Simple Evolutionary Model

- We first need an (simplified/oversimplified) evolution model:
  - Only substitution and indel
  - Two mutations do not overlap
    - Guarantee that all evolutionary information but the order is represented by the alignment.
  - Along the path of evolution, p: unchanged, q: substitution, r: indel. (p+q+r=1)
- ```
  ATGCA-TGTA   (S)
  |||  |  ||  |
  ATGTACTG-A   (T)
  ```
  ```

# Probability of the Alignment

- Under this model, the probability of the alignment is:
  - $p^7 * q * r^2$
- We want to maximize this probability
  - $\log (p^7 * q * r^2) = 7 \log p + \log q + 2 \log r$
- Let match = log p, mismatch = log q, indel = log r. We get a scoring scheme.
- Maximizing the score is equivalent to maximizing the probability of evolutionary history.

# Simple Score

- This is sufficient to compare the alignments of the same two sequences.
- Problems
  - Always negative
  - Local alignment becomes meaningless
  - Repeating the alignment twice make the score lower.
- Useless in comparison of two alignments of different pairs of sequences.
- Question: Can these be solved by adding a minus sign?
- Next let us make this score better.

Which solution is better? – different instances.

# PURPOSE 2

# Likelihood Ratio

- Model 1 (homology): the alignment A between S and T reflects evolutionary history.
- Model 2 (random): the alignment A between S and T is merely a random event.
- We want to examine the likelihood ratio
  - Pr(alignment|homology) / Pr(alignment|random)
- If it's much bigger than 1 (such as 100000), it's evidence towards model one being the truth.
- If it's much below 1 (such as 0.00001), it's evidence towards model two being the truth.

# Log likelihood Ratio

- Assume for the homology and random models, we have established the probabilities for each column type:
  - Match: p and p'
  - Substitution: q and q'
  - Indel: r and r'
- For the following alignment,
  - ```
    ATGCA-TGTA  (S)
    |||  |  ||  |
    ATGTACTG-A  (T)
    ```
  - Pr(alignment|homology) / Pr(alignment | random) = $(p/p')^7 * (q/q') * (r/r')^2$
- We usually take a logarithm. The sore becomes
  - 7* log (p/p') + log (q/q') + 2 * log (r/r').
- If this is very positive, then homology model explains the alignment better than random model. And vice versa.
    ```

# This is a Better Scoring Scheme

- Score scheme prefers column types happens more often in the homology model than in the random model. Usually,
  - p>p', therefore a matching column has a positive score
  - q<q', therefore a mismatching column has a negative score
  - r <r', therefore an indel column has a negative score.
- Avoids the problems we had when only probabilities (not the ratio) were used.
  - Putting two positive alignments together increase the homology chance.
  - Can compare two alignments with different lengths.
- This is indeed the scoring scheme we have seen and have used in practice (in BLAST etc.)

# Statistics

- The probability values used in the homology and random models may be obtained by simple counting their frequencies in some "real" alignments and "random" alignments, respectively.
- Often, the statistics is only approximate and does not need to be precise. In particular, the "random" model often uses some (over)-simplied values.
  - For example: Pr(indel)= 0.2, Pr(match) = 1/20 * 0.8, Pr(mismatch) = 19/20 * 0.8.

# Substitution Matrix

- The non-indel columns can be further refined to have different scores for different pairs of letters.
- For each pair of letters a and b, assume the probability of seeing (a,b) in a column is p(a,b) for the homology model, and is q(a,b) for the random model.
- Then substitution score is then log (p(a,b)/q(a,b)).
- This is called a substitution matrix.
- Let us assume that q(a,b)=p(a)*p(b). Here p(a) is the frequency of letter a in the sequences. Note that this is an (over)-simplification. But it provides "good enough" values in practice.

# Substitution Matrix

- The substitution matrix is particularly important when aligning protein sequences because
  - there are 20 amino acids
  - some of them share significant similarities
  - protein alignments have fewer matching columns.

# Alignment of Protein Sequences

```
Conserved domain database 22426:
KOG4652, HORMA domain [Chromatin structure and dynamics]

Conserved domain length = 324 residues, 100% aligned

CT46          15  VFPNKISTEHQSLVLVKRLLAVSVSCITYLRGIFPECAYGTRYLD--DLCVKILREDKNCPG--STQLVKWMLGC
                    PN +  E QSL  + RLL V++S I   RGIFPE  +  RY+D   L + +LR      G   + L K +
KOG4652       1   TLPNGLENEKQSLEFMTRLLYVAISTILRERGIFPEEYFKDRYVDGNLLVMTLLRRQDAPEGRLVSWLEKGV---

CT46          85  YDALQKKYLRMVVLAVYTNPEDPQTISECYQFKFKYTNNGPLMDFISKN------QSNESSMLSTD-TKKASILL
                  +DA+++K L+ + L V T  EDP+ I E Y F F Y   G +   I+        ++ E S LS D T++    L
KOG4652       73  HDAIRQKLLKKLSL-VITESEDPEDI-EVYIFSFVYDEEGSVSARINYGINGQSSKAFELSQLSMDDTRRQFAKL

CT46         154 IRKIYILMQNLGPLPNDVCLTMKLFYYDEVTPPDYQPPGFKDGDCEGVIFEGEPMYLNVGEVSTPFHIFKVKVTT
                  IRK++I  Q L PLP    +     YY E  PPDYQP GFKD           P  +N+G VSTP H   VKV
KOG4652      146 IRKLHICTQLLEPLPQ-GLILSMRLYYTERVPPDYQPEGFKDSTRAFYTLPVNPEQINIGAVSTPHHKGFVKVL-

CT46         229 ERERMENIDSTILSPKQIKTPFQKILRDKDVEDEQEHYTSDDLDIETKMEEQEKNPASSELEEPSLVCEEDEIMR
                                                    SD  D   K E
KOG4652      219 ----------------------------------------SDATDSMEKAER----------------------T

CT46         304 SKESPDLSISHSQVEQLVNKTSELDMSESKTRSGKVFQNKMANGNQPVKSSKENRKRSQHESGR---IVLHHFDS
                  K S D       V+Q +NK+ E D S S+ ++  +  N + N   PV  S+E+  +SQ   G          D
KOG4652      232 DKISDDP-FDLILVQQELNKSEEADKSFSQEKTTSITPNVLGNPLVPVDQSEEDLLKSQDSPGTGRCSCECGLDV

CT46         376 SSQESVPKRRKFSEPKEHI
                 S Q SVPK RK     EH
KOG4652      306 SKQASVPKTRKSCRKTEHG


Homology between CT46 and MGC26710 hypothetical protein

Identities = 136/249 (54%), with conservative changes = 180/249 (72%)

CT46          1   MATAQLQR-----TPMSALVFPNKISTEHQSLVLVKRLLAVSVSCITYLRGIFPECAYGTRYLDDLCVKILREDK
                  MATAQL             VFP++I+ EH+SL +VK+L A S+SCITYLRG+FPE +YG R+LDDL +KILREDK
MGC26710      1   MATAQLSHCITIHKASKETVFPSQITNEHESLKMVKKLFATSISCITYLRGLFPESSYGERHLDDLSLKILREDK

CT46         71   NCPGSTQLVKWMLGCYDALQKKYLRMVVLAVYTNPEDPQTISECYQFKFKYTNNGPLMDF--ISKNQSNESSMLS
                   CPGS  +++W+ GC+DAL+K+YLRM VL +YT+P   + ++E YQFKFKYT  G  MDF   S + S ES   +
MGC26710     76   KCPGSLHIIRWIQGCFDALEKRYLRMAVLTLYTDPMGSEKVTEMYQFKFKYTKEGATMDFDSHSSSTSFESGTNN

CT46         144 TDTKKASILLIRKIYILMQNLGPLPNDVCLTMKLFYYDEVTPPDYQPPGFKDG-DCEGVIFEGEPMYLNVGEVST
                  D KKAS+LLIRK+YILMQ+L PLPN+V LTMKL YY+ VTP DYQP GFK+G +   ++F+ EP+ + VG VST
MGC26710     151 EDIKKASVLLIRKLYILMQDLEPLPNNVVLTMKLHYYNAVTPHDYQPLGFKEGVNSHFLLFDKEPINVQVGFVST

CT46         218 PFHIFKVKVTTERERMENIDSTIL  241
                  FH  KVKV TE  ++ ++++ +
MGC26710     226 GFHSMKVKVMTEATKVIDLENNLF  249
```

Notice the blocks with no indel.

# BLOSUM 62

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

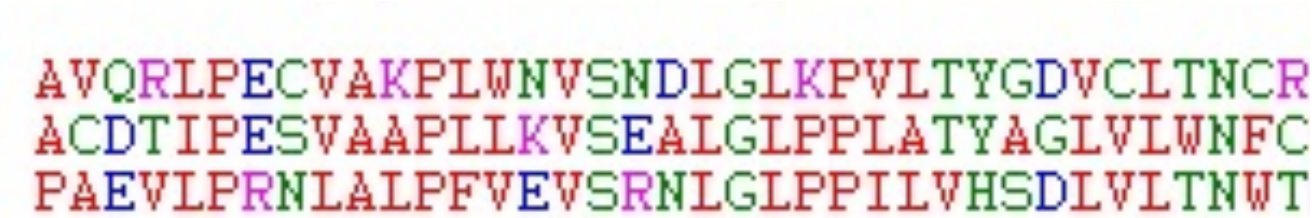
Non-standard letters

B= D or N
Z= E or Q
X= any
* = translation stop

The most used amino acid substitution matrix. Let's study how this is constructed.

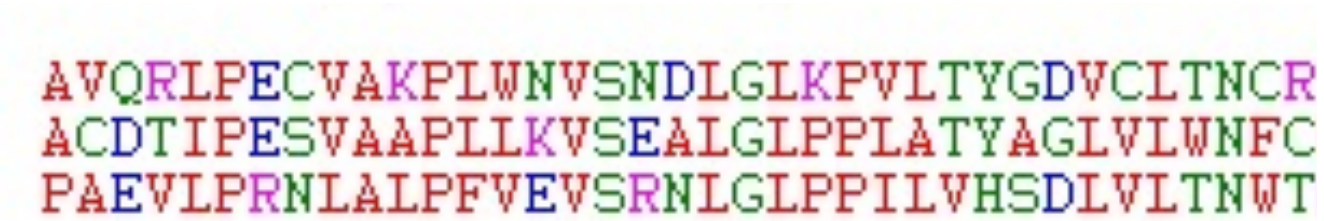
# Basic idea

- Conserved regions from multiple sources are aligned into blocks.




- The identity level is high therefore we know they are homologues without a score matrix.
- There was such a database consisting those blocks.
  - Henikoff S, Henikoff JG. Automated assembly of protein blocks for database searching. Nucleic Acids Res. 1991 Dec 11;19(23):6565-72. doi: 10.1093/nar/19.23.6565. PMID: 1754394; PMCID: PMC329220.
  - Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A*. 1992;89(22):10915-10919. doi:10.1073/pnas.89.22.10915

# Frequency of AA pairs




- 37 columns, each column 3 pairs. In total 111 pairs.
- For example, the pair I-L occurs 3 times; the pair L-L occurs 13 times

$$P_{IL} = \frac{3}{111}, P_{LL} = \frac{13}{111}$$

- Total amino acid 111 (a coincident).

$$P_I = \frac{2}{111}, P_L = \frac{21}{111}$$

- We can then use log likelihood ratio to calculate scores.
- But we should correctly distinguish the counting when two letters are the same or different.

# Blosum

$$score(x,y) = 2\log_2 \frac{P_{xy}}{2P_xP_y}, \text{ if } x \neq y$$

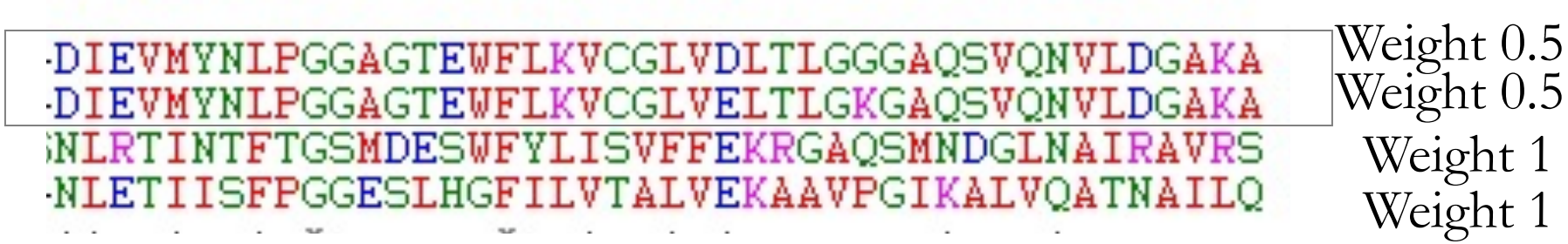
$$score(x,x) = 2\log_2 \frac{P_{xx}}{P_xP_x}$$

In BLOSUM matrices these values are rounded to the nearest integer.

# Deal with Sample Bias

- Some protein families are more well studied so they are over-represented in the database.
- Such bias is caused by the studies, not reflecting what's going on during evolution.
- To remove this bias in statistics, those "redundant" proteins are classified together before BLOSUM calculation.

# BLOSUM 62




- The sequences that are 62% or above similarity are grouped together and given total weight 1.
- This way, the AA pairs are counted between groups that are 62% similar or below.
- The lower this number is, the better is the matrix suitable to distant homology search.
- The original BLOSUM paper found out 62 is best at the time the paper was prepared.

# A quick review

- Where does a score come from?
- Homology model: Two sequences are homologous using the alignment/evolutionary history.
- Random model: Two sequences are irrelevant.
- Which model better explains the alignment?
- Log likelihood ratio
- If greater than zero, supports Homolog, else, supports Random.

- Statistics should be done carefully to avoid oversampling.

Does the alignment indicate a homology?

# PURPOSE 3

# Significance

- Consider that BLAST matches a query sequence with all database sequences, and return the highest scoring local alignments.
- The best local alignment score is 100. Does it mean good or bad?
- The answer depends on the score scheme you use so we need to standardize it for effective communication.
- Intuitively, we would ask "can this happen randomly"?
- This is formalized as the p-value.

# P-value

- You want to prove that a coin is biased on its two sides.
- Null hypothesis $H_0$: the coin is fair
- Alternative hypothesis: the coin is biased
- Experiment: You draw the coin 20 times, and found 14 heads and 6 tails.
- Under the null hypothesis, how extreme are these values?
- $\Pr(\#\text{head} \geq 14 \text{ or } \#\text{tail} \geq 14 \mid H_0) = 0.1154$
- This is called the P-value.

# P-value

- In statistical hypothesis testing, the **P-value** is the probability of obtaining a result at least as extreme as the one that was actually observed, given that the null hypothesis is true.
- A small P-value **rejects** the null hypothesis. So, we choose to believe the alternative hypothesis.
- In the coin example, 0.1154 is certainly not good enough.
- But what if we draw 40 times, and found 28 heads and 12 tails?
- Now P-value is 0.0115.
- This is how scientists show the effectiveness of a treatment through clinicals. (null hypothesis: it is not effective).

# P-value

- A small P-value **rejects** the null hypothesis. So, we choose to believe the alternative hypothesis.
  - **This does not mean that the alternative hypothesis is surely correct. It is just the null hypothesis is unlikely to be correct. It is just a way to communicate the significance.**
  - E.g. "If the coin is fair, then we should not have seen such an imbalance (28:12) in 40 flips."
  - E.g. "If the two sequences are irrelevant, then we should not have seen the alignment with such a high score." (But they may be related for reasons other than homology – e.g. convergence evolution.)
- In practice an arbitrary threshold 0.05 is often used, for no apparent reason.

# P-value for Local Alignment

- Null hypothesis: the query S and the database T are irrelevant.
- Now we observe a high scoring local alignment with score x. This is an extreme case.
- What is the P-value?
- How to compute the P-value?
- From now on, S and T are long.

# Statistics of Ungapped Local Alignment

- For ungapped alignment (no indel), each local ungapped alignment is called an HSP (High Scoring Pair) in the BLAST program.
- For an individual HSP, the alignment score is the sum of n identical independent variables. So the score is a binomial distribution.
- But we have a lot of such ungapped alignments, and we only output the best one (with the highest alignment score).
- For P-value, we need to check how rare this best alignment has a high score.
- The maxima of many identical independent variables is a so-called extreme value distribution.

# Statistics of Ungapped Local Alignment

- The following two papers studied the distribution of ungapped local alignment (HSPs) scores via both E-values and P-values.
  - Karlin, S. & Altschul, S.F. (1990) "Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes." Proc. Natl. Acad. Sci. USA 87:2264-2268.
  - Dembo, A., Karlin, S. & Zeitouni, O. (1994) "Limit distribution of maximal non-aligned two-sequence segmental score." Ann. Prob. 22:2022-2039.

# E-value

**Theorem 1**: In the limit of sufficiently large sequence lengths $m$ and $n$, the statistics of HSP scores are characterized by two parameters, $K$ and $\lambda$. Most simply, the expected number of HSPs with score at least $x$ is given by the formula

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

- This is called the E-value of the alignment.
- The parameters $K$ and $\lambda$ can be somehow determined from the scoring matrix. (Details not given here).
- Thus, we just convert the alignment score $x$ to E-value. If it is very small, then random sequences will not often produce HSPs with score $\geq x$.
  - Likely caused by homology.
  - Not many false positives if we treat every local alignment with score $\geq x$ as homologs.

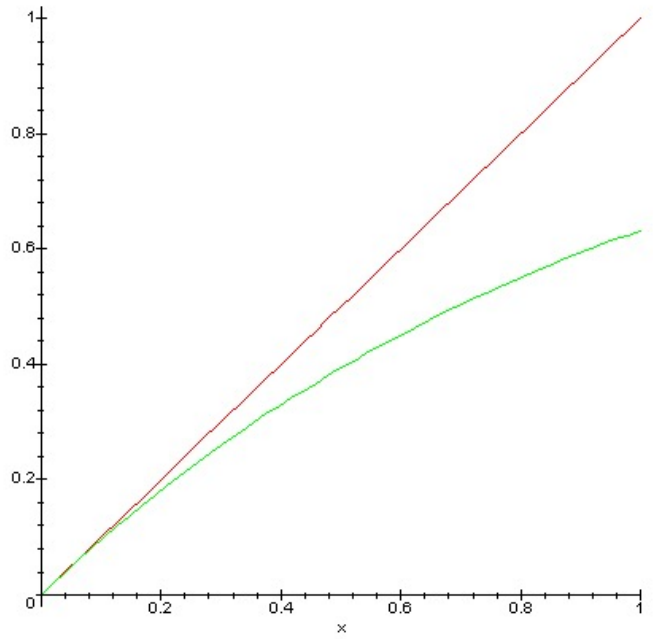
# P-value

- **Theorem 2** (proved in those two papers): The number of random HSPs with score $\geq S$ is described by a Poisson distribution. Therefore the chance of finding zero HSP with score $\geq S$ is $e^{-E}$
- Therefore the P-value is $1- e^{-E}$
- When E value is very small, P-value and E-value are almost identical.

# E-value v.s. P-value

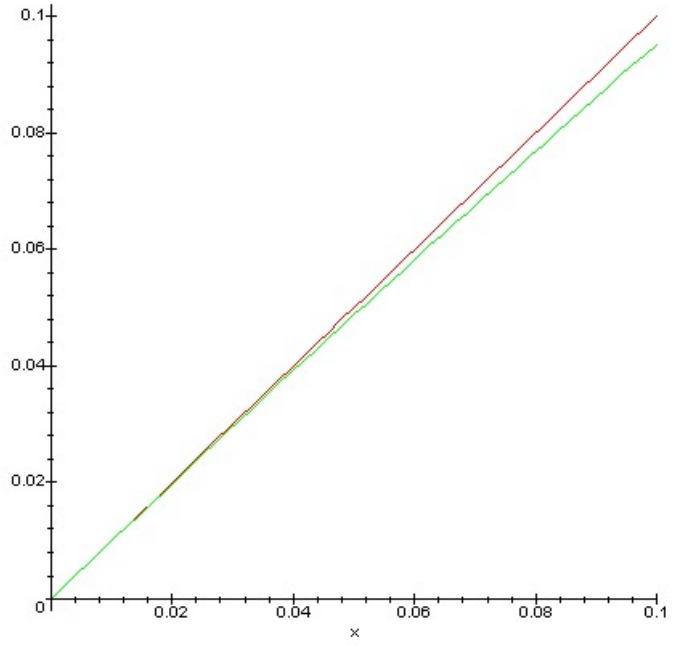
- BLAST chose to use E-value.

P-value




E-value

P-value




E-value

# Advantage Over "Raw" Score

- The E-value and P-value have their physical meanings.
  - E.g. If S and T are random, we expect to see on average 0.01 HSP with such a high score.
- The alignment score, however, largely depends on the scoring matrix one chose.
- Saying that "the alignment score is 100" is like saying "the length is 100", meters? feet?

# The statistics of gapped alignments

- The statistics developed above have a solid theoretical foundation only for local alignments that are not permitted to have gaps. However, many computational experiments and some analytic results strongly suggest that the same theory applies as well to gapped alignments.
- But we need to know how to compute K and lambda.
- Question: How can we estimate the two parameters K and lambda?

# Wrap up

- Alignment score purpose 1: to compare alignments of the same two sequences.
- Purpose 2: to compare alignments for different pairs of sequences.
- Purpose 3: to determine if the two sequences are homologous.
- Alignment score is log likelihood ratio: log (likelihood of homology / likelihood of random)
- BLOSUM
- Assess local alignment: E-value, P-value