EE550
Computational Biology

Week 7 Course Notes

Instructor: Bilge Karaçalı, PhD

Topics

- Searching sequence databases
  - Gene and protein identification by sequence similarity
    - Sequence database queries
    - Performance evaluation
  - Sequence alignment statistics

Sequence Similarity

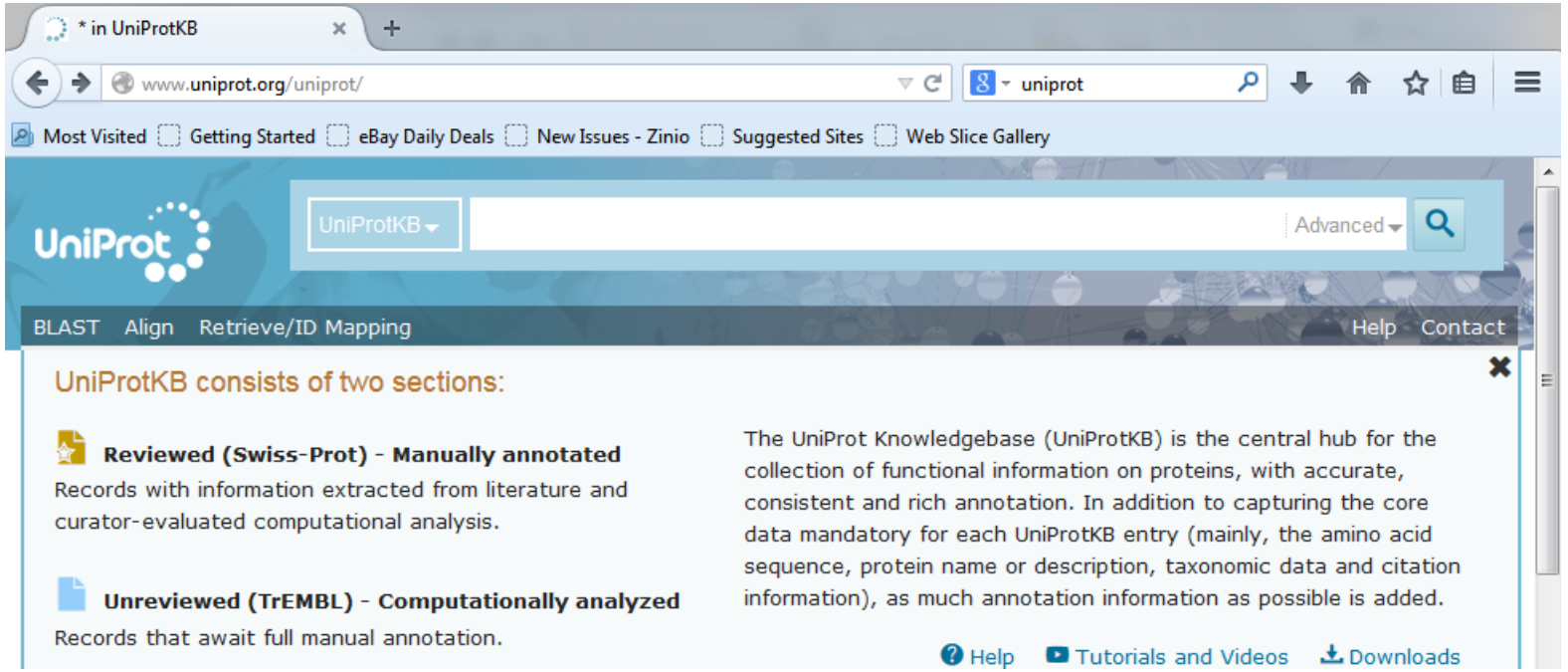
- Sequence similarity suggests **common origin** and **similar function**
  - Organisms start with a common ancestor and evolve along different paths
    - Each organism acquires different sets of mutations
    - Common ancestry is established by the persisting sequence traits
  - As the sequences of genes diverge, the sequences of the corresponding proteins diverge as well
    - The divergence of the proteins implies selective control on the evolutionary mechanism
      - Mutations that reduce the fitness are strongly selected against
    - This control ensures that the necessary functionality of the protein is preserved all through the evolutionary time-frame
- Comparing a new sequence to a database of known sequences allows
  - Estimating its evolutionary relationships to known molecules of known organisms
  - Predicting its function (as well as protein structure)

Sequence Databases

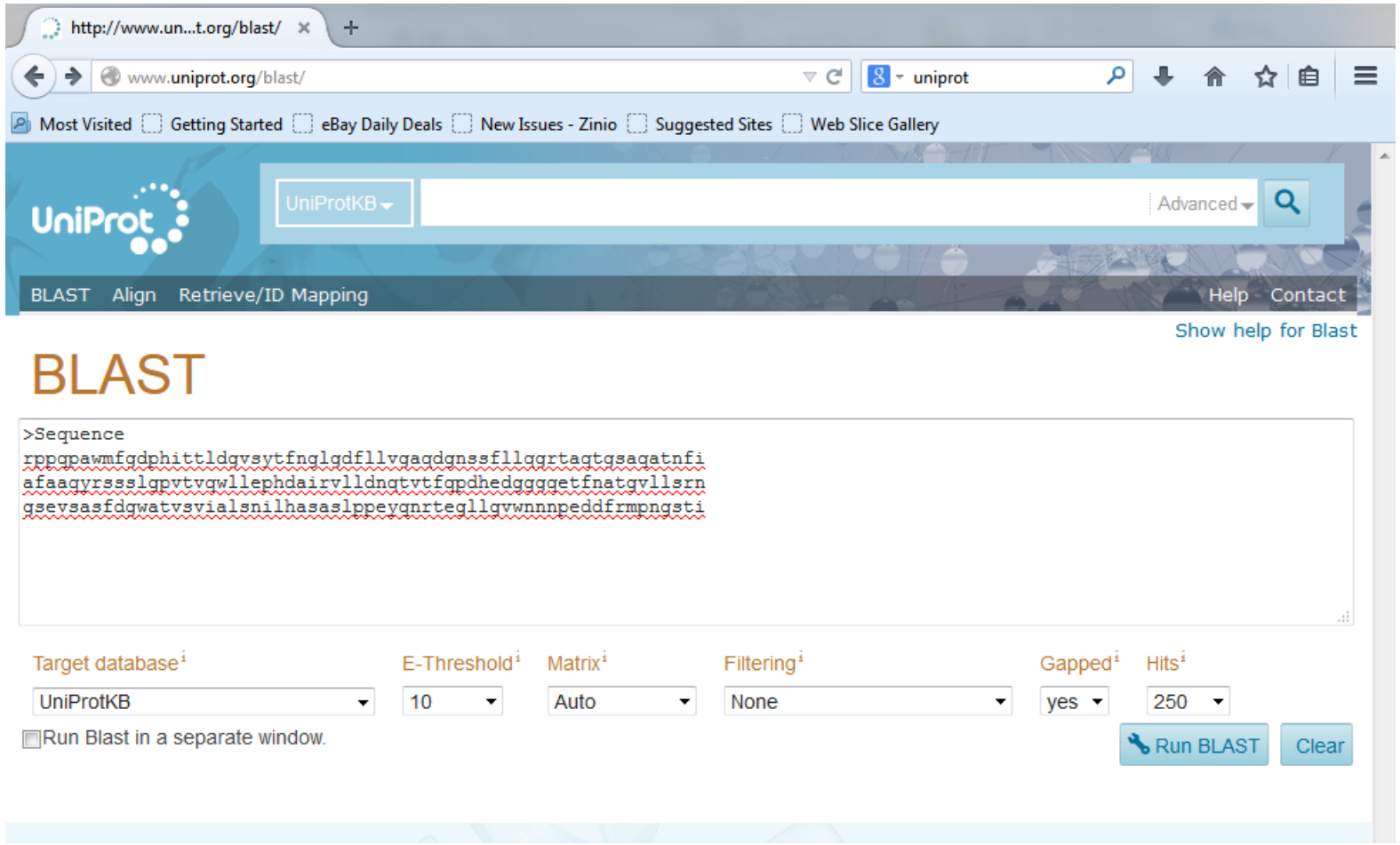
- Several online databases allow access to large collections of genetic and protein sequences
  - Genes:
    - EMBL
    - GenBank
    - DDBJ
    - …
  - Proteins:
    - PIR
    - MIPS
    - UniProt
    - NCBI Protein Database
    - GenPept
    - …

Sequence Database Queries

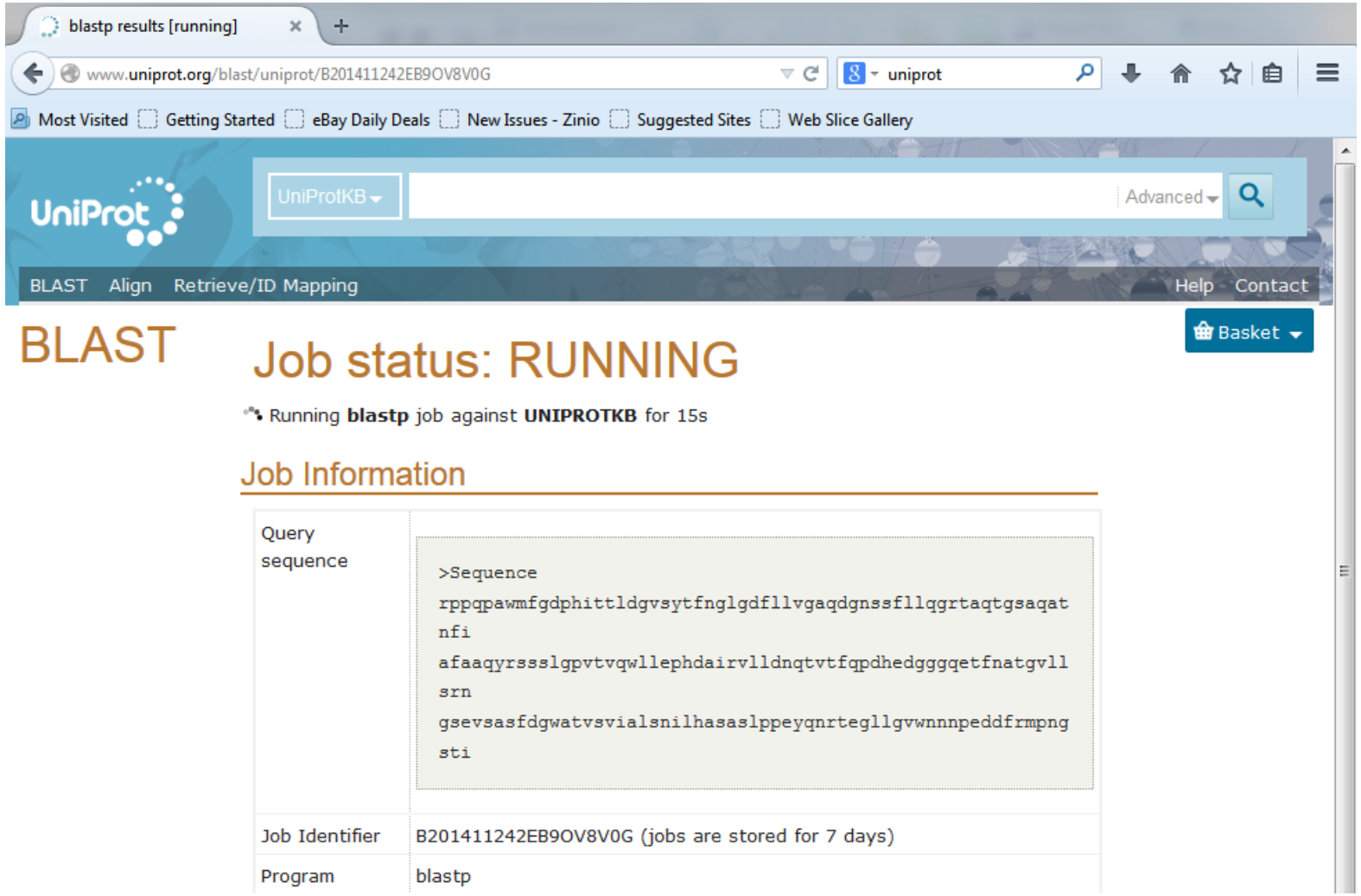
- **Task:** Given a sequence fragment
  `rppqpawmfgdphittldgvsytfnglgdfllvgaqdgnssfllqgrtaqtgsaqatnfi`
  `afaaqyrssslgpvtvqwllephdairvlldnqtvtfqpdhedgggqetfnatgvllsrn`
  `gsevsasfdgwatvsvialsnilhasaslppeyqnrtegllgvwnnnpeddfrmpngsti`
  - Query the UniProt sequence database for proteins with similar sequences
  - Evaluate the list of proteins with highest similarity scores



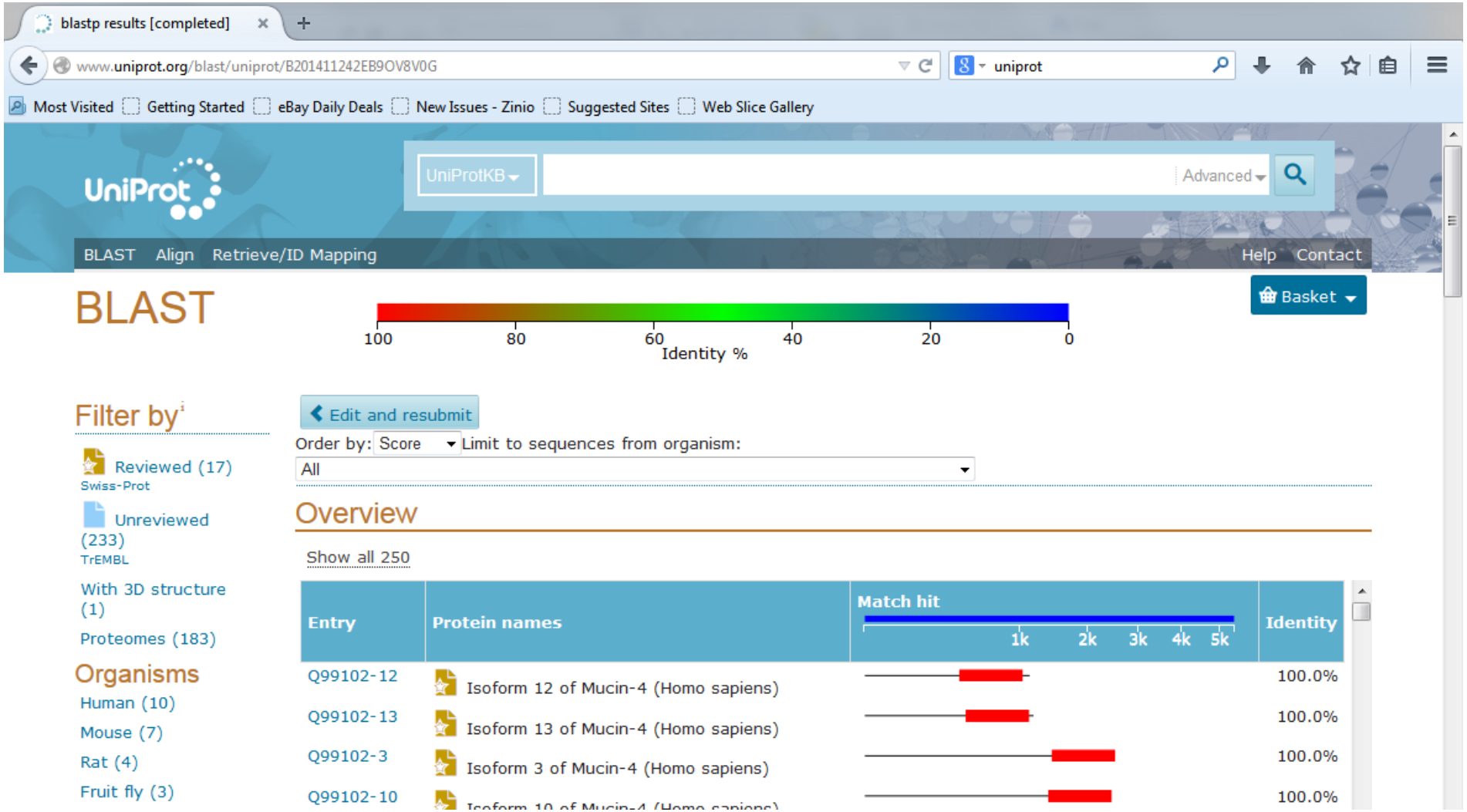
Sequence Database Queries



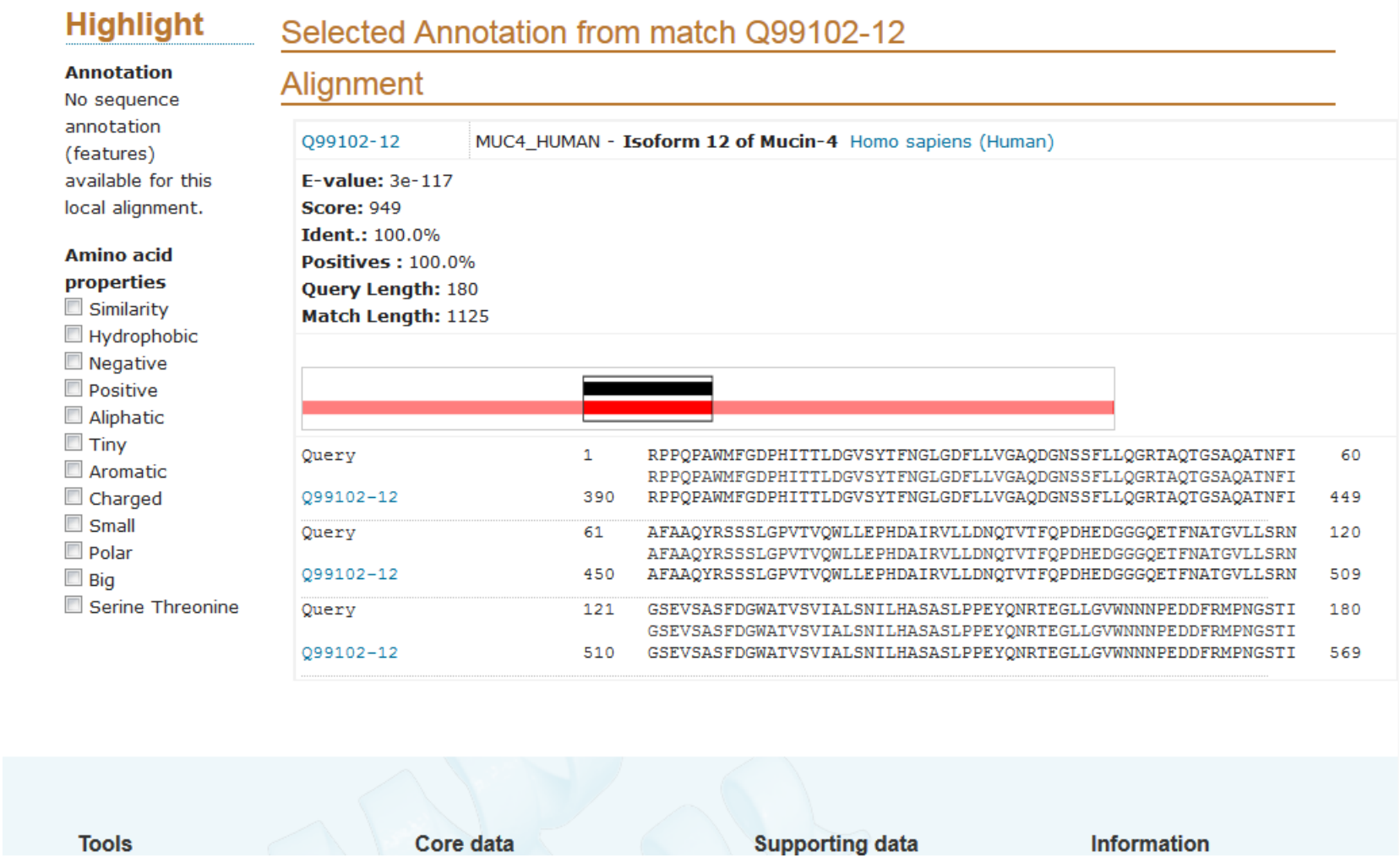
Sequence Database Queries



Sequence Database Queries



Sequence Database Queries



Sequence Database Queries

- UniProt sequence query:
  - The given sequence fragment is aligned to all sequences in the database
    - Pairwise sequence alignments by BLAST
  - The database entries with the highest alignment scores are returned in descending order
    - The most similar sequence is listed first
    - The similarity score monotonically decreases down the list
  - Identification of sequence properties, function, and evolutionary relationships is carried out based on the set of most similar proteins

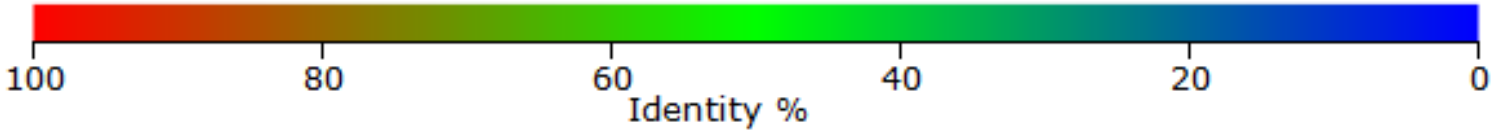
  Q: How similar does a protein need to be so that it can safely be recognized?

Sequence Database Queries

- UniProt sequence query (continued):
  - Several fields for every identified protein are returned
    - Familiar fields:
      - Accession number
      - Entry name
      - Protein name
      - Organism
    - Novel fields:
      - Local alignment
      - Identity
      - Positives
      - Score
      - E-value
      - Query length
      - Match length

Sequence Database Queries

- UniProt sequence query (continued):
  - Local alignment
    - Provides a graphical display of how the given sequence fragment was aligned with that of the entry
    - The gaps embedded into the sequence fragment are not shown
      - Does not mean that there have not been any gaps
    - Similarity is represented by a color code
      - Bright red: fully of nearly identical
      - Brown-green: quite similar
      - Blue: not similar at all



Sequence Database Queries

- UniProt sequence query (continued):
  - Identity
    - Varies between 0 and 100
    - Denotes the percentage of amino acids that are identical in the overlap
      - The overlap constitutes the aligned regions of both sequences
        - » The query sequence
        - » The database sequence
      - A score of a 100 implies all amino acids in the overlap are the same
      - On the other hand, a score near 0 implies none of the amino acids match
        - » Very hard to see (Q: Why?)
    - The gaps and substitutions are treated the same
  - Positives
    - Percentage of sites for which the similarity score is positive
      - log-odds similarity

Sequence Database Queries

- UniProt sequence query (continued):
  - Score
    - Denotes the alignment score; maximized by the BLAST algorithm between the query sequence and the database sequence
    - Uses the log-odds scoring matrix for amino acid replacements and the specified gap penalty function
  - The relative likelihood is measured by

$$RL = \log\big(R(\boldsymbol{A}, \boldsymbol{B})\big) = \sum_i S_{\boldsymbol{A}_i, \boldsymbol{B}_i}$$

where *S* is the specified scoring matrix (e.g., PAM250, BLOSUM80, …)

  - The overall gap penalty is measured by

$$GP = \sum_j f\big(\ell_j\big)$$

where j indexes the gaps with lengths ℓ_j, and f denotes the chosen gap penalty function (e.g., linear, affine, …)

  - The overall alignment score is then

$$\text{Score} = RL - GP$$

Sequence Database Queries

- UniProt sequence query (continued):
  - E value
    - Denotes the number of sequences with which the observed alignment quality can be observed by pure chance
      - A random sequence of comparable length and composition is provided to the search engine over the same database
      - The distribution of alignment scores are observed

      Q: How many sequences in the database will be aligned to a similar random sequence with a score no less than the observed score?

      A: E value!!
    - Requires a model of the probability distribution of alignment scores with all sequences in the database
    - Takes values ranging from nearly 0 to tens or hundreds
      - Low values (significantly smaller than 1) suggest that the alignment could not have been observed by chance so that there must be something worthy of attention
      - High values (around 1 or higher) suggest whatever the observed relationship, it might very well be due to chance

Sequence Database Queries

- UniProt sequence query (continued):
  - E value (continued)
    - Consider the sequence

      PAIRWISEALIGNMENTTOOLSHAVEVARIOUSINHERENTISSUESNOTTHELEASTTHATDIFFERENTPROGRAMSOFTENRETURNDIFFERENTRESULTS
    - BLAST on this sequence in the UniProt database returns several hits with different E values
    - The high E values attest to the poor quality of the alignments
    - This gives us grounds to reject the alignments

Sequence Database Queries

Entry	Alignment overview	Info	Status
Query: B201411242EBD9QU4QE			
A0A073IM30	A0A073IM30_9RHOB - Helicase - Sulfitobacter do... - View alignment 	E-value: 3.2e1 Score: 75 Ident.: 29.0%	
V4AEI7	V4AEI7_LOTGI - Uncharacterized protein - Lottia gigantea ... - View alignment 	E-value: 4e1 Score: 70 Ident.: 33.0%	
A0A067QRM0	A0A067QRM0_ZOONE - Uncharacterized protein - Zootermopsis nev... - View alignment 	E-value: 4.1e1 Score: 74 Ident.: 25.0%	
A4BJL3	A4BJL3_9GAMM - Putative acetyltransferase - Reinekea blanden... - View alignment 	E-value: 4.3e1 Score: 73 Ident.: 25.0%	
A3U822	A3U822_CROAH - Cell division protein FtsZ - Croceibacter atl... - View alignment 	E-value: 1.1e2 Score: 71 Ident.: 32.0%	
Q804X3	Q804X3_CHICK - Coagulation factor VIII - Gallus gallus (C... - View alignment 	E-value: 1.1e2 Score: 71 Ident.: 32.0%	
F1NPT2	F1NPT2_CHICK - Uncharacterized protein - Gallus gallus (C... - View alignment	E-value: 1.1e2 Score: 71	

Sequence Database Queries

Selected Annotation from match A0A073IM30

Alignment

A0A073IM30 A0A073IM30_9RHOB - **Helicase** Sulfitobacter donghicola DSW-25 = KCTC 12864 = JCM 14565

E-value: 3.2e1
Score: 75
Ident.: 29.0%
Positives : 46.0%
Query Length: 106
Match Length: 982

```
Query          5    WISEALIGNMENTTOOLSHAVEV---------ARIOUSINHERENTISSUESNOTTHELE    55
                    W +EAL+G+ ++ T  LSH V++         A    +  HE     +  E+     ELE
A0A073IM30     468  WAAEALMGHQDHQTYPLSHWVDLHLKIANDVAAGTSGNTEHELWQQKAGMEARRVMQELE   527

Query          56   ASTTHATDI    64
                    A   + TD+
A0A073IM30     528  AEAGYGTDL   536
```

Performance Evaluation of Query Algorithms

- Protein sequence databases are queried primarily to establish functional familial relationships
  - A given sequence is compared to those in a sequence database
  - The database sequences with the highest alignment similarity to the sequence in question are listed in a descending order
  - The queried sequence is then presumed to **belong to the functional family that is most represented** among the statistically significant query hits
    - A query hit is a database sequence with a significantly high sequence alignment score
      - Very low E value
- The success of the querying algorithm is measured in its ability to group
  - the correct familial proteins at the top of the hit list with high statistical significance and
  - everything else in the bottom with no statistical significance

Performance Evaluation of Query Algorithms

- Performance evaluation is critical to predict the success rates of alternative methods
  - Alternative alignment methods
  - Different parameter choices for the chosen alignment method
- **Ideally**, performance evaluation would be conducted on unseen-before data (i.e. real life testing)
  - The operation would be carried out for a newly sequenced protein
    - Alignment against a database
    - Ranking of the statistically significant hits in terms of the alignment scores
    - Prediction of the familial relationships
    - Testing of these predictions in conventional wet-lab experiments
  - The average performance for several such proteins would estimate the performance of the employed algorithms

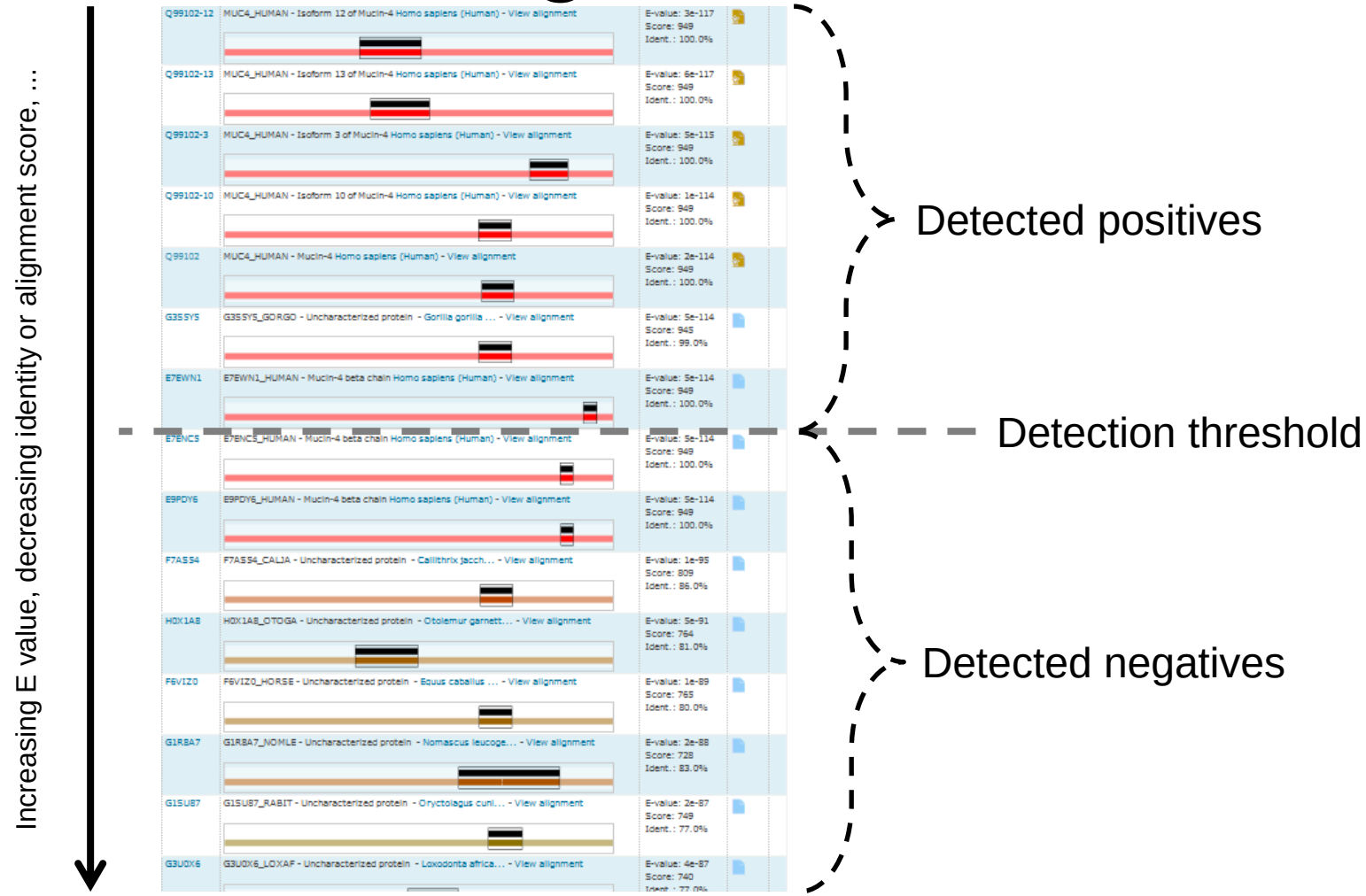
Performance Evaluation of Query Algorithms

- But the ideal procedure is neither feasible nor viable:
  - Requiring the whole wet-lab verification process for each unseen-before sequence for performance evaluation is extremely costly
  - Using hard-collected data on unseen-before proteins to test the performance of prediction algorithms defeats their purpose
    - The whole reason for employing such algorithms is to be able to predict the functional and familial properties of newly-sequenced proteins without the wet-lab procedures
- Instead, cross-validation techniques from the statistical learning literature are used for performance evaluation
  - A functional protein group is identified for the purpose
    - Transcription factors
    - Antigen-binding proteins
    - Kinases
    - …
  - The prediction procedure is carried out for a small subset of the protein group by removing them from the working database
  - The recognition performance is evaluated in terms of how many members of the protein group in consideration are identified beyond a statistical significance threshold

Performance Evaluation of Query Algorithms

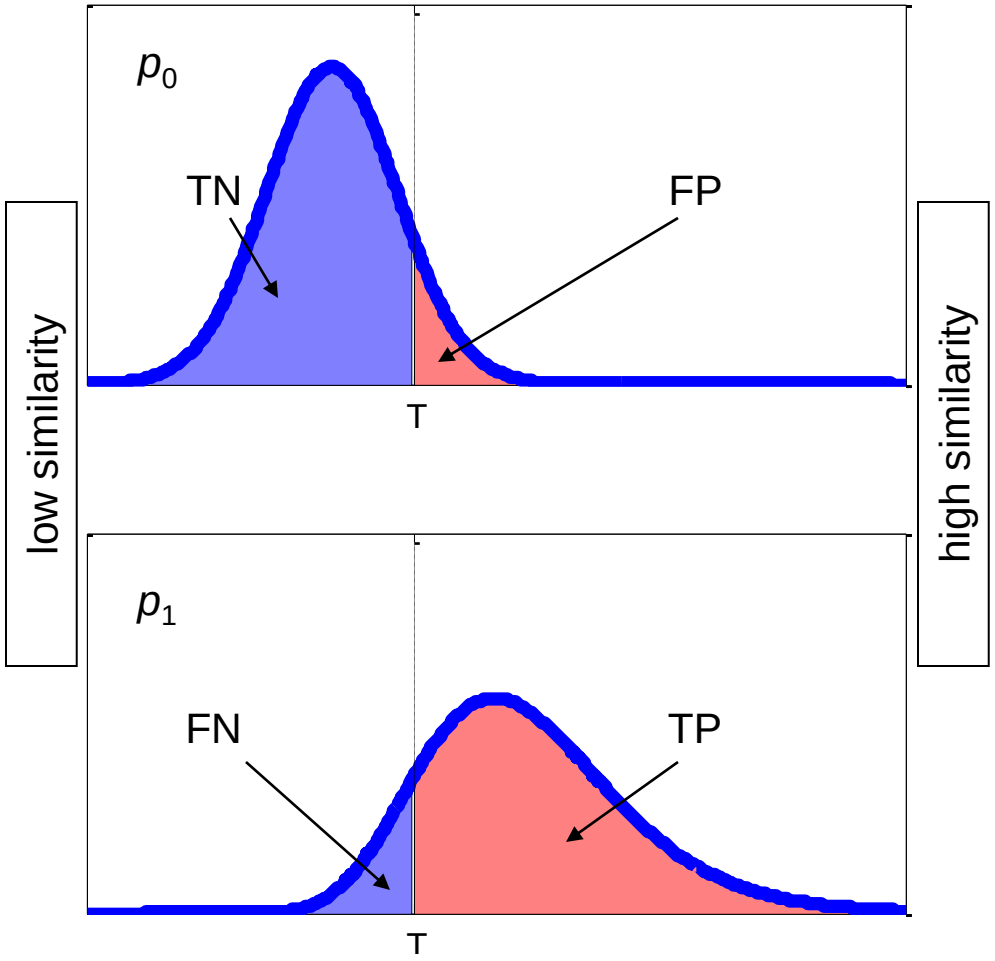
- Procedure:
  - A randomly selected member of a protein group of interest is queried against a database that consists of
    - the sequences of the remaining members of the protein group, C_1, and
    - the sequences of all the other proteins in the original database, C_0
  - The relevant statistics (alignment score, identity, E value, …) are compared against varying threshold levels for detection
  - The number of sequences in C_1 and C_0 above or below a given detection threshold are counted
    - The proteins with statistics satisfying the threshold are “detected”
    - Conversely, the proteins with statistics failing to satisfy the threshold are “not detected”
  - Performance of the prediction algorithm is measured in terms of the respective fractions of C_1 and C_0 that are “detected,” respectively, correctly and incorrectly

Performance Evaluation of Query Algorithms



Performance Evaluation of Query Algorithms

- The values of the statistic observed on C_0 and C_1 constitute the distributions p_0 and p_1
- The integrals of p_0 and p_1 computed over the intervals $(-\infty, T)$ and (T, ∞) determine the basic performance markers
  - True negatives: $\int_{-\infty}^{T} p_0(t)dt$
  - False positives: $\int_{T}^{\infty} p_0(t)dt$
  - False negatives: $\int_{-\infty}^{T} p_1(t)dt$
  - True positives: $\int_{T}^{\infty} p_1(t)dt$
- In actual applications, the distributions are replaced by histograms
  - TN, FP, FN, and TP become protein counts and not fractions



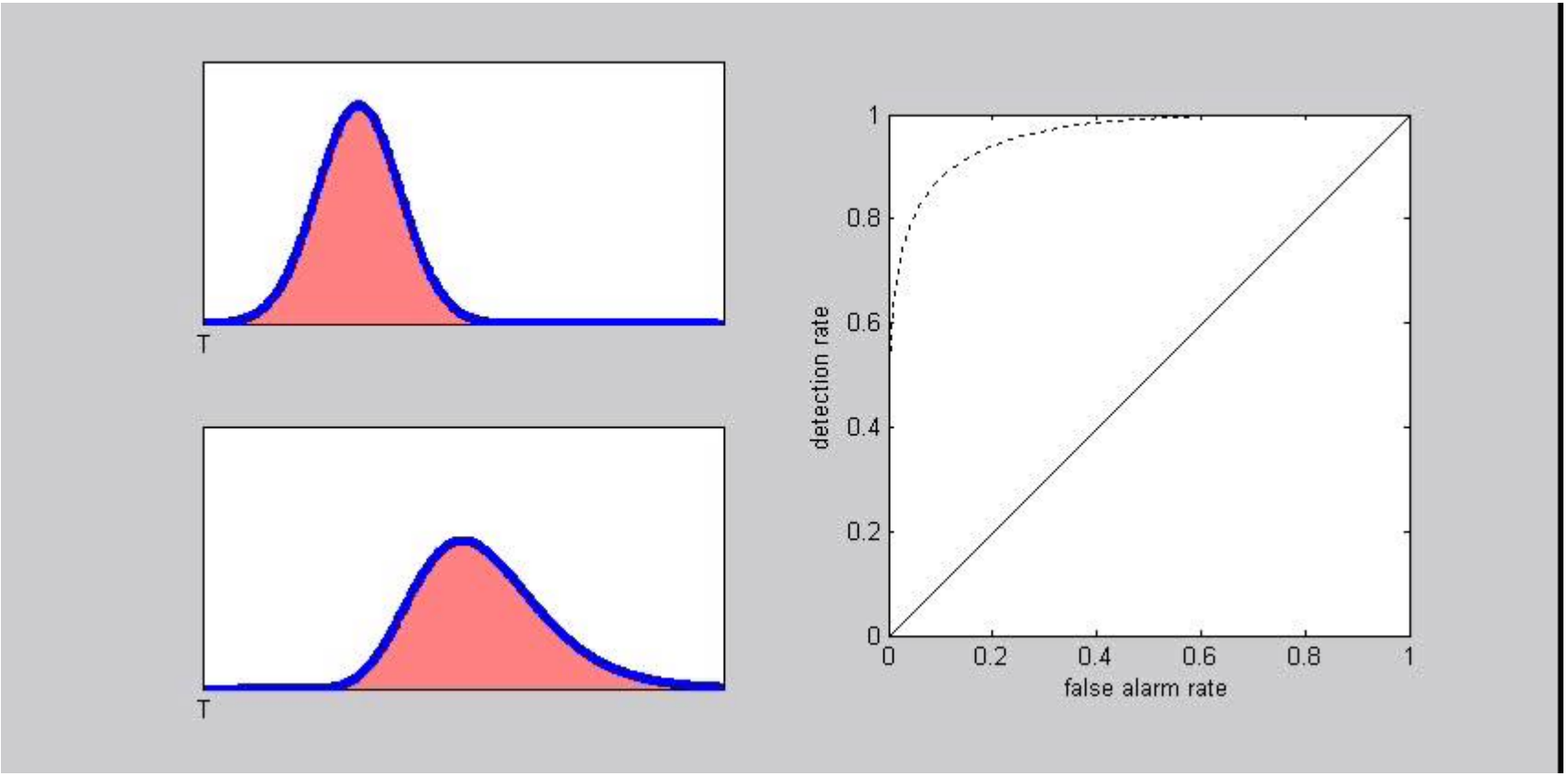
Performance Evaluation of Query Algorithms

- Higher order performance measures are computed from the basic quantities of TN, FP, FN, and TP
  - Detection rate = TP/(TP+FN) {sensitivity}
  - False detection rate = FP/(TN+FP) {false alarm rate}
  - Specificity = TP/(TP+FP) {selectivity}
  - F-measure = 2·sensitivity·specificity/(sensitivity+specificity)
- The variation of the **detection rate** with respect to the **false alarm rate** is called

the receiver operating characteristics

of the detection rule
  - The area under this curve provides an average measure of performance irrespective of a threshold

Performance Evaluation of Query Algorithms



More on Alignment Statistics

- Given any sequence fragment, querying it against a sequence database will always return a **ranked list of entries**
  - The ranking is typically returned in the order of decreasing alignment score starting from the entry with the highest score
- The real question is whether any of the obtained alignments carries any **significance** at all
  - Databases are queried in order to establish functional and familial relationships between the sequence at hand to the sequences in the database
  - The results obtained from the query are believable only if they are supported by a **statistical significance analysis**
  - Otherwise, the obtained good scores may very well have been **accidental**, and hence, **meaningless**
- The statistical significance of the results is determined against random queries of comparable nature
  - The observed results are significant if the probability of observing them is *really* low

Pairwise Alignment Statistics

- Consider the pairwise alignment of two nucleotide sequences of length N
  - The probability *p* that a given site is occupied by the same nucleotide in both sequences by pure chance is

$$p = \pi_{\mathrm{A}}^2 + \pi_{\mathrm{T}}^2 + \pi_{\mathrm{G}}^2 + \pi_{\mathrm{C}}^2$$

    where π_A, π_T, π_G, and π_C denote the prior probabilities of the corresponding nucleotides
  - The expected number of sites occupied by the same nucleotide, whatever that nucleotide may be, is then $p \cdot N$
    - Note, however, that the probability of having **all** sites in two sequences of length N match is p^N, assuming independence of sites
  - The number m of sites occupied by the same nucleotides in both sequences by pure chance follows a binomial distribution with the probability mass function

$$p_B(m) = \binom{N}{m} p^m (1-p)^{N-m}$$

  - The probability P of observing greater than or equal to m matching sites is

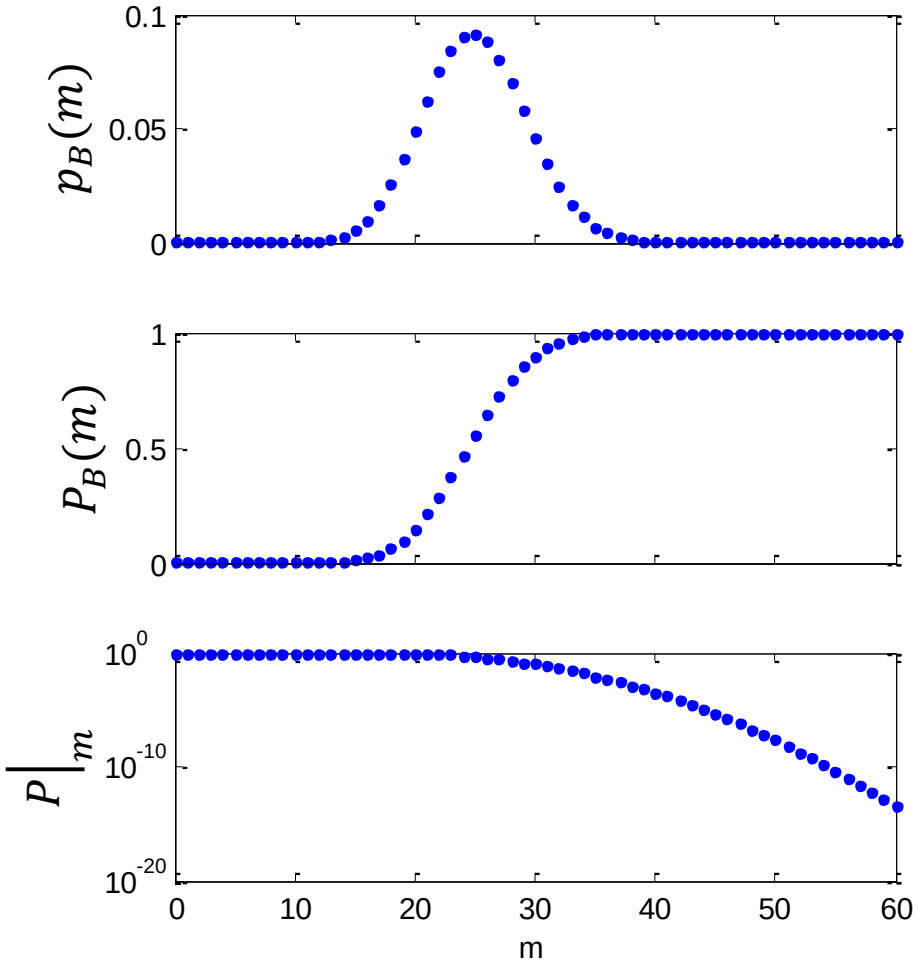
$$P = \sum_{m'=m}^{N} p_B(m')$$

  - The smaller the P value, the less likely to observe m matching sites by pure chance

Pairwise Alignment Statistics

- Example:
  - Let
    - $\pi_{\mathrm{A}} = \pi_{\mathrm{T}} = \pi_{\mathrm{G}} = \pi_{\mathrm{C}} = 1/4$
    - $N = 100$
  - The probability distribution then becomes

    $$p_B(m) = \binom{100}{m}(0.25)^m(0.75)^{100-m}$$
  - The P values associated with observing various m numbers of matching sites can be obtained as
    - $P|_{m=25} = 5.38 \cdot 10^{-1}$
    - $P|_{m=35} = 1.64 \cdot 10^{-2}$
    - $P|_{m=45} = 1.09 \cdot 10^{-5}$



Query Alignment Statistics

- When a whole set of sequence alignments are evaluated for statistical significance, the probability structure of the experiment changes
  - Instead of one observation, we will need to sort out several observations simultaneously
- The question then becomes whether any of the observed similarity scores are higher than the expected maximum in a random case
  - In the random case, a comparable but random sequence is queried against the dataset
  - If the distribution of the resulting alignment scores can be obtained, then the distribution of the maximal scores can be modeled as well
  - The expected maximum can then be obtained as the mean of the distribution of the maximal scores
  - This maximal distribution can also be used to compute the P values associated with observing a certain maximal score in chance experiments

  → Extreme Value Distributions

Extreme Value Distributions

- Extreme value distributions govern the statistical behavior of extreme events
  - Maxima
  - Minima
- Note that extreme events are also random variables
  - Let X be a random variable, and $\{X_i\}$, $i = 1, \ldots, n$, denote a random collection of n independent and identically distributed random variables with the same distribution as X
  - Define M as the maximum of the collection $\{X_i\}$

$$M = \max_i \{X_i\}$$

  - Note that
    - M is a random variable as well, and
    - the distribution of M is an **extreme value distribution**

Extreme Value Distributions

- The probability distribution of the maxima

$$M = \max_i \{X_i\}$$

$$\begin{aligned} \Rightarrow F_M(x) &= \Pr\{M \le x\} \\ &= \Pr\{X_1 \le x, X_2 \le x, \ldots, X_n \le x\} \\ &= \Pr\{X_1 \le x\} \cdot \Pr\{X_2 \le x\} \cdot \ldots \cdot \Pr\{X_n \le x\} \\ &= (\Pr\{X \le x\})^n \\ &= \big(F_X(x)\big)^n \end{aligned}$$

$$\Rightarrow f_M(x) = \frac{d}{dx} F_M(x) = n\big(F_X(x)\big)^{n-1} f_X(x)$$

  - For discrete distributions, increments at integer x produce the corresponding probability mass functions

Extreme Value Distributions

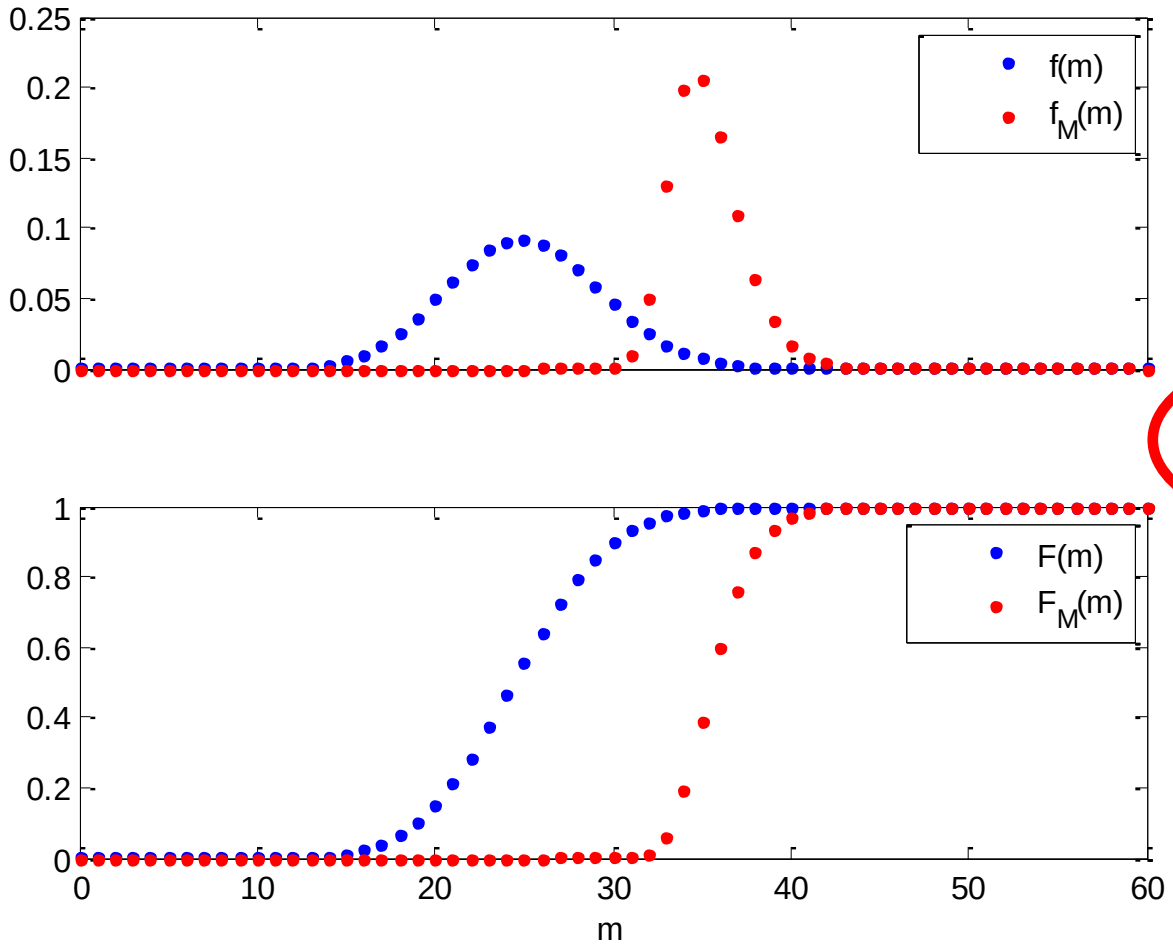
- Consider the case where a nucleotide sequence of length N is queried in a database of n sequences, each of length N
- The objective is to compute the extreme value distribution governing the maximum alignment score between the dataset sequences and a random sequence of length N
- Procedure:
  - Given the binomial probability distribution of the pairwise match score $f(m)$ for sequence pairs of length N
  - Compute the associated cumulative distribution function $F(m)$
  - Compute the cumulative distribution function $F_M(m)$ by

$$F_M(m) = \big(F(m)\big)^n$$

and the extreme value distribution's probability mass function by

$$f_M(m) = F_M(m) - F_M(m-1)$$

Extreme Value Distributions



$n = 100$

Remarks

- The extreme value distribution for the maximal alignment scores on random sequences estimates the number of random hits that would be included for a given threshold
  - The E values provided by the UniProt query system corresponds to the expected number of random hits with the **same or better similarity score** in the **same database**
- Note that in actuality, the extreme value distribution is quite difficult to obtain (numerically or in closed form)
  - Sequence databases are not random collection of arbitrary sequences
    - These sequences are the products of millions of years of selection
  - The alignment scores from one sequence to the next are not necessarily independent from one another
    - The sequences in the database usually belong to distinct sequence families
  - A viable approach is to sample the distribution using alignments with random sequences of varying length and composition, and then to generalize to suitable extreme value distribution models

Summary

- Sequence databases provide online utilities that allow submitting queries with novel sequences
- These queries determine the most similar sequences in the database to the queried sequence
- A common functional or familial grouping among the most similar database sequences is suggestive of similar functionality and lineage
- The degree at which one should trust the identified hits lies in the level of statistical significance
  - Usually provided by the E values in query result tables