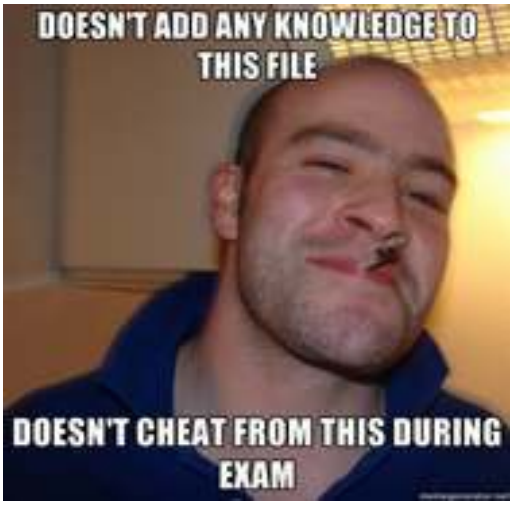
Na, legyen az a taktika, mint az előzőnél: tutijó megoldás, lehet-hogy-jó megoldás, hol találtam
[Geri]: csak nekem gyanús, hogy a 20. kérdéstől kezdődően nem adták le az anyagot, hanem helyette sok sok filogenetika volt?
[Norbi]: Én rákerestem mindegyik diában a 7.től kezdve kb 5 kulcsszóra azokból a feladatokból, de sehol sem volt találat. (kivéve a p-value-t)



1. What is the difference between retrieval and similarity search in databases?

09_Similarity_search_BLAST page 5
• Given a query and a database, find the entry in the database that is most similar to the query in terms of a numerical similarity measure (distance, similarity score, etc.) - **Similarity search**
• In contrast: **retrieval** looks for an exact match to the query.
• Is John on the paylist? **Retrieval**: Yes/No, based on exact matching. **Similarity search**: His brother Joe Brown is. So we can classify John into the Brown family, based on approximate matching.

2. What are the two types of heuristics used in sequence similarity searching? Describe both of them briefly! 09_Similarity_search_BLAST page 14-22

I. Computational heuristics:
a) exact matching is fast,
b) search space can be easily reduced.
- Word based methods:
  • Exact word matching is fast, use it whenever you can.
  • Word composition vectors are a possibility but: word (n-gram) dimensionality grows fast with word size
  • Longer words are more informative but few of the possible words occur in a given sequence, and most of them only once…
  • We can make use of the longer words if we look only for **shared (common) words**. These are fewer…
  • Potentially shared words are all the words of two sequences… (for two 100-amino acid sequences, how many words are in total?

• Word indexing methods, such as hash tables are extremely fast. But you need to build an index (extra overhead).

-Search space reduction:

• Most sequences in a dquerybase are not similar to a given query. It is easy to construct filters that throw away uninteresting sequences.

• Threshold based filtering is important, but has drawbacks (you can throw away important hits)

• **Use fast methods to filter out most of the dbase entries, and use increasingly expensive methods only for the important ones**.

• Typical filter 1: Keep only sequences that contain n common words with the query.

• Typical filter 2: The shared words should be near. Or, they should be in the right sequence.

II. Biological heuristics

include a) local similarities are dense, b) similar regions are near each other, c) low complexity sequences excluded, etc.

- Search space reduction:

• If you absolutely need dynamic programming, search only the vicinity of the diagonal

-Short conserved sequences:

• Sequence related by evolution always contain conserved regions that are highly similar, contain rows of identical sequences… Omit those that do not….

• Danger: difficult to define, what is a conserved region (and what is a good threshold)

-Repetitive “low complexity” regions:

• Biological sequences often contain repetitive of “biased composition”, “low complexity” regions.

• Such sequences contain a biased composition of n-gram words, they are thus less complex (can be described with fewer words).

• They give background noise since e.g. AAAAAA would match all sequences containing A-s.

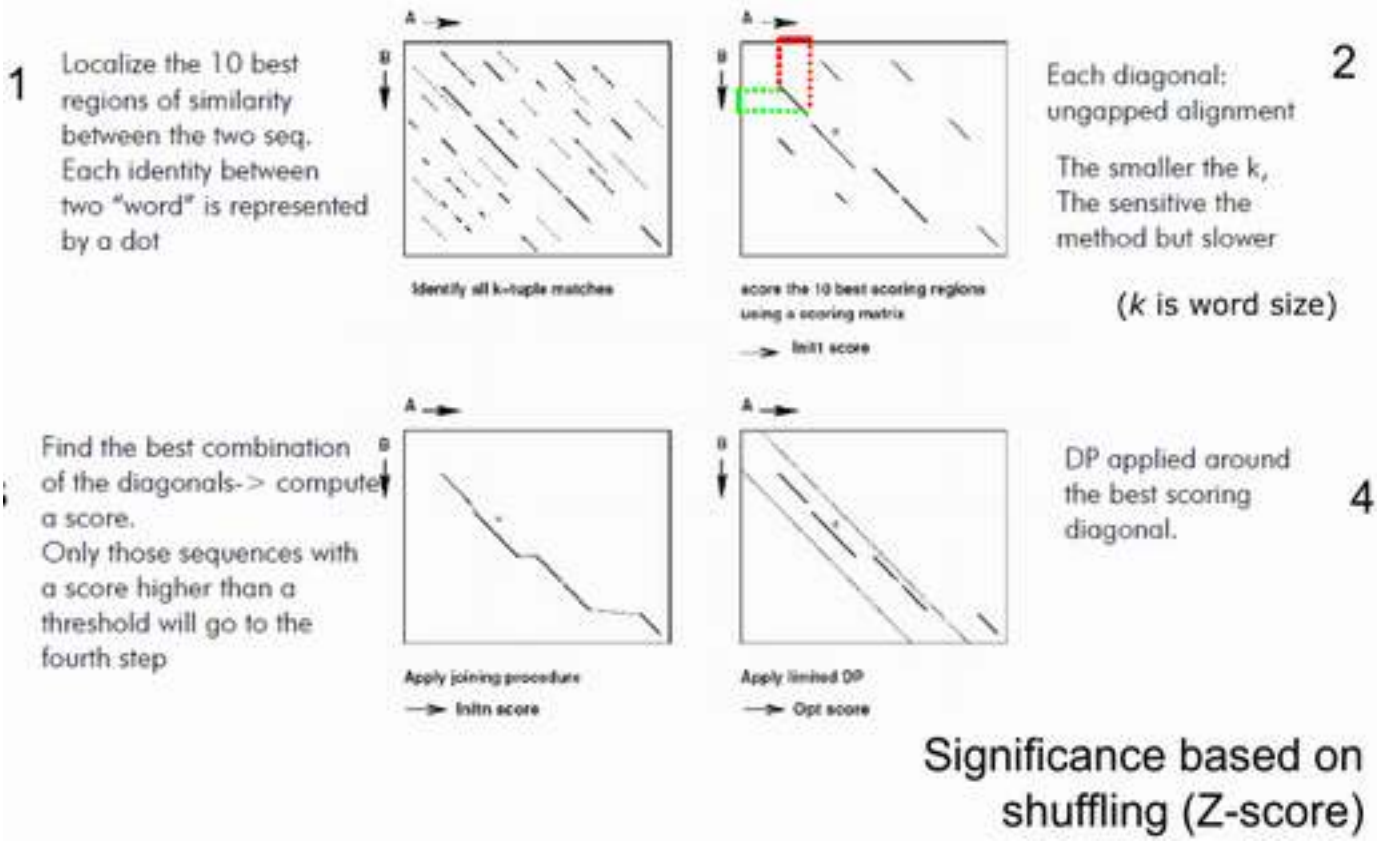
• Such low-complexity regions should be excluded from the comparisons. SEG program of John Wootton masks them with X-es, before comparison.

3. List the 4 steps of the original FASTA algorithm! What is the problem with the algorithm? 09_Similarity_search_BLAST page 25-26

Steps:

1. Localize the 10 best regions of similarity between the two sequences. Each identity between two “words” is represented by a dot.
2. Each diagonal: ungapped alignment. The smaller k means more sensitivity, but it slows down the algorithm. (k is the word size)

3. Find the best combination of the diagonals -> compute a score. Only those sequences with a higher score than a threshold will go to the next step.
4. Dynamic programming is applied around the best scoring diagonal.



Problems:
• Used word identities to calculate initial scores (not too sensitive)
• Score significance was based on the Z-score (random shuffling is not too fast)

4. Describe the main steps of BLAST algorithm!

Wikipédia és http://www.cs.ucsb.edu/~ambuj/Courses/bioinformatics/sequence-dbsearch.pdf
1. Make words of length 3 (for proteins) or 11 (for DNA) from the query sequence using a sliding frame

MEFPGLGSLGTSEPLPQFVDPALVSS
MEF
EFP
FPG
PGL

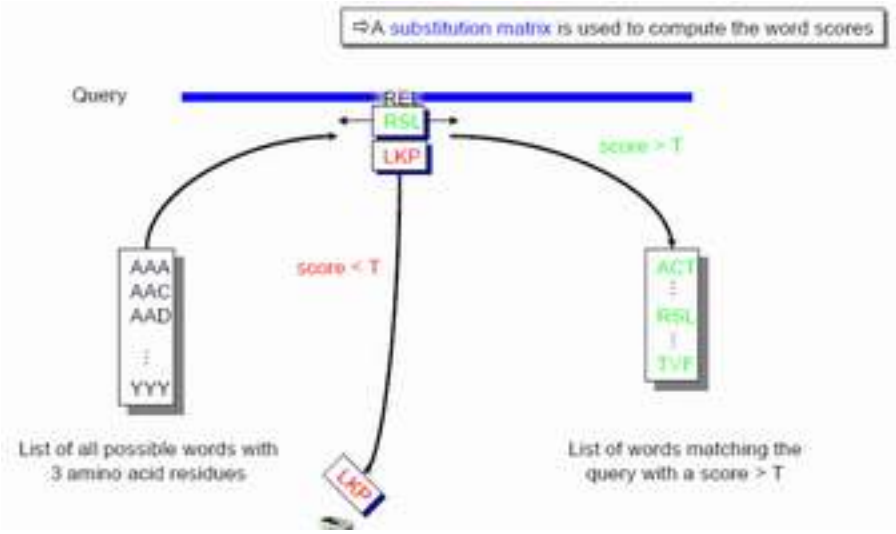
2. List and score the possible matching words.
3. Select a neighborhood word score threshold(T) so that only most significant hits are kept. Organize the high-scoring words into an efficient search tree.
4. Scan database sequence for an exact match to high-scoring words. Each match is a seed for an ungapped alignment.

5. Extend matching words to the left and right using ungapped alignments, as long as score increases or stays same. This is an **HSP** (high scoring pair).
6. Using a cutoff score S, keep only those HSPs that have a score at least S.
7. Determine the statistical significance of each remaining HSP.

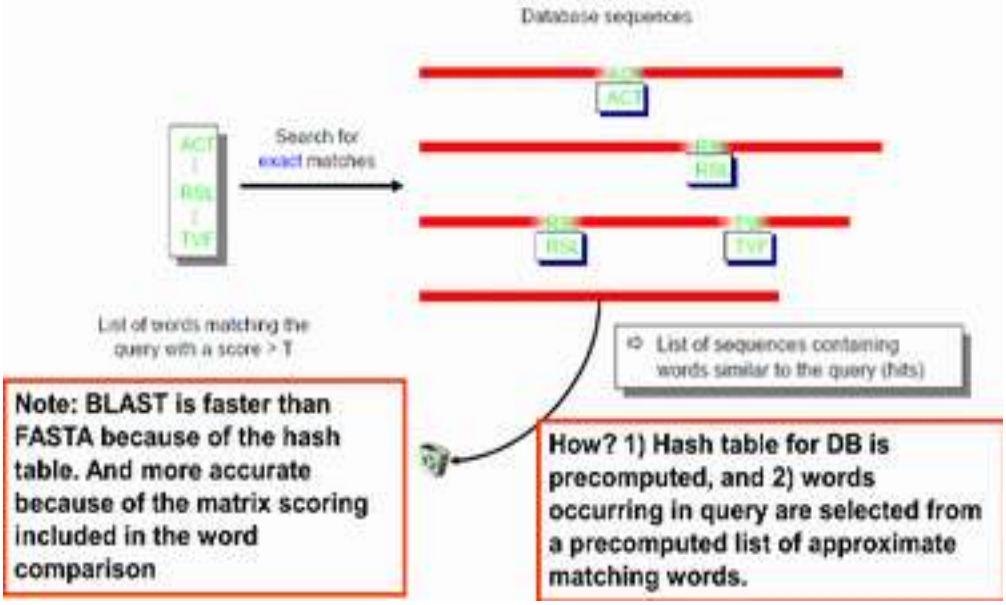
+optional: Make two or more HSP regions into a longer alignment.
Report every match whose expect score is lower than threshold parameter E

09_Similarity_search_BLAST page 27-32

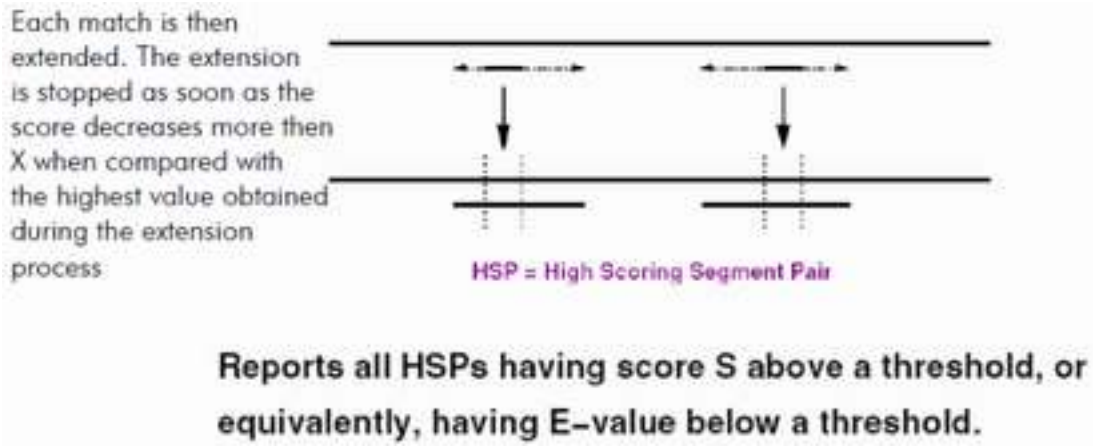
- 1. Creating a list of similar words
  This is an extended word list that allows mismatches in a search for exactly matching words!!

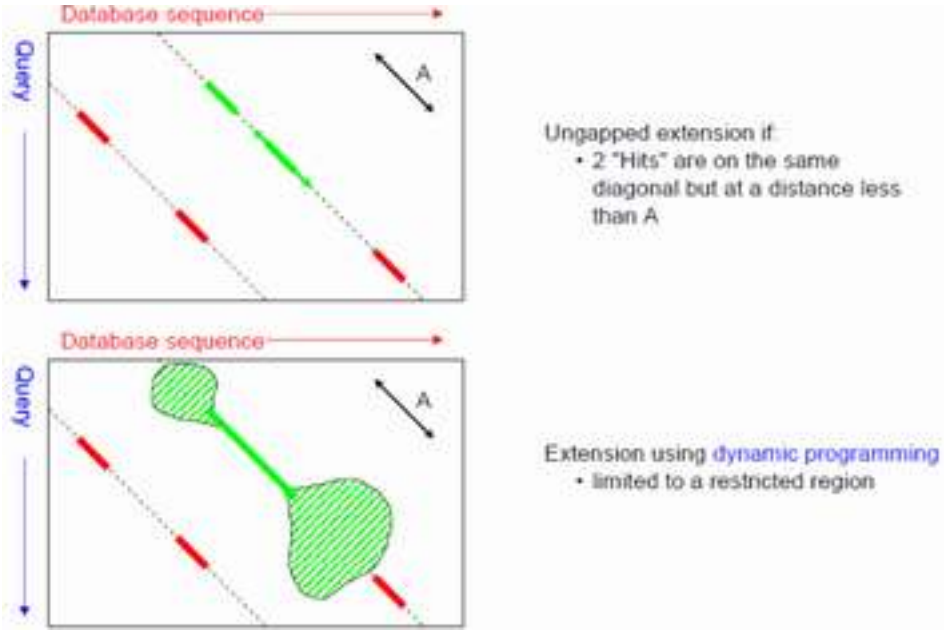


- 2. Eliminating sequences without word hits



- 3. Extension of hits
  For each word match (<hit>), extend ungapped alignment in both directions. Stop when S decreases by more than X from the highest value reached by S.





This step generates ungapped HSPs.

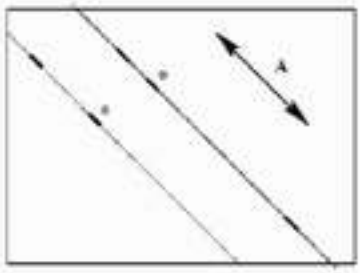
- 4. Gapped extension of HSPs having score above a threshold Sg

- 1+2+3+4. step:

First step: as with BLAST1, generate lists of words scoring more than T with words of the query.

Second step: generation of hits: identify all word matches in DB sequences

Third step: extension of hits: requires a second hit on the same diagonal at a distance of less than A.



Additional step:
Gapped extension of the hits slower-> therefore: requirement of a second hits on the diagonal. (hits not joined by ungapped extensions could be part of the same gapped alignmnet)

This step generates ungapped HSPs

Fourth step: gapped extension of HSPs having score above a threshold S_g

09_Similarity_search_2a: page 10:

Ranking (e.g. Blast) in general:

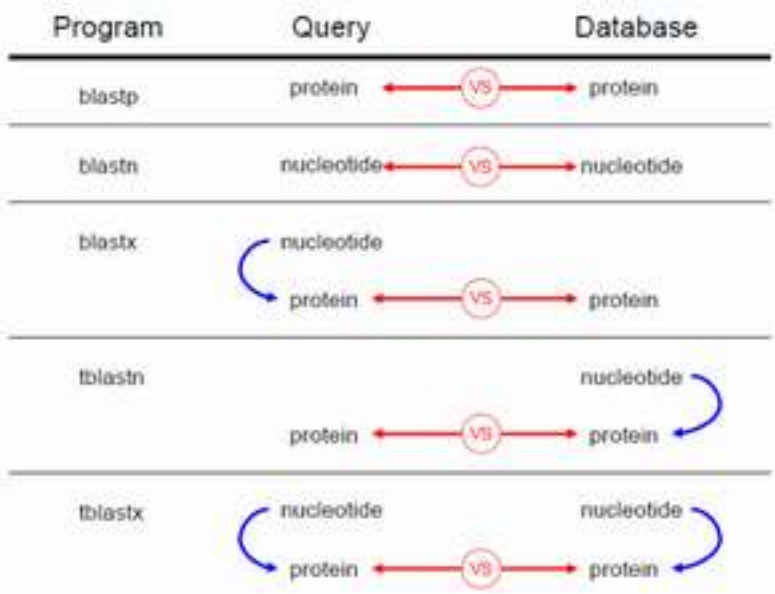
- Calculate a distance function between a query and each object in a database
- Rank the objects according to the distance (or similarity)
- Turn ranking into a classifier: apply a threshold. Similarities better than threshold are positives (TP, FP), below threshold are negatives (TN, FN).

5. Write down the BLAST representation of this sequence: ACATGACCTCTC. Use words of 3 character!

ACA, CAT, ATG, TGA, GAC, ACC, CCT, CTC, TCT, CTC

6. List 3 BLAST programs with different query types and explain what they do!

09_Similarity_search_BLAST page 42-43



blastp: compares a protein sequence with a protein database

-to find something about the function of your protein, compare it with other proteins in the database

-identify common regions between proteins, or collect related proteins

tblastn: compares protein sequence with nucleotide database

-to discover new genes encoding proteins, use tblastn to compare your protein with DNA sequences translated into their six possible reading frames; map a protein to genomic DNA

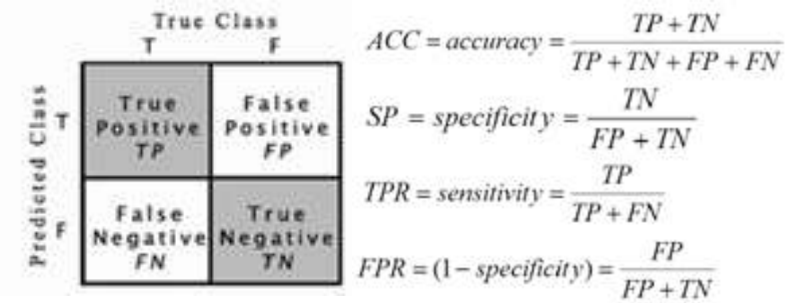
blastn: compares nucleotide sequence with nucleotide db

7. What are the advantages of BLAT against BLAST?

- speed (no queues, response in seconds) at the price of lesser homology depth
- the ability to submit a long list of simultaneous queries in fasta format
- five convenient output sort options
- a direct link into the UCSC browser
- alignment block details in natural genomic order
- an option to launch the alignment later as part of a custom track

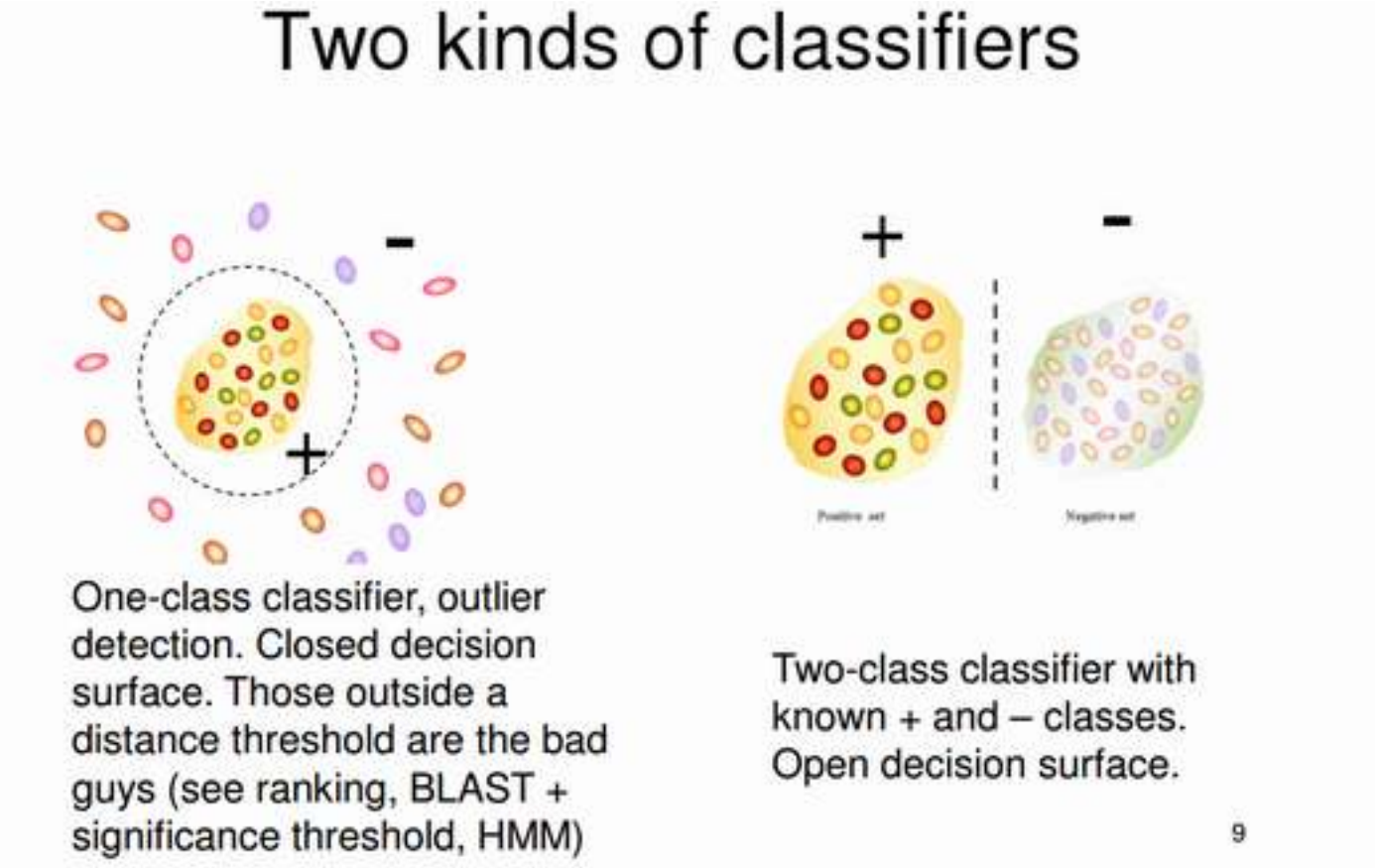
8. What is binary classification? What is confusion matrix?

Classification is an algorithmic procedure for assigning an input object to a known category. Binary classification is based on two classes, e.g. true or false. Binary decisions can be described by a confusion matrix.



09_Similarity_search_2a: 6.page

9. Draw a sketch of one-class and two-class classification. Explain what the difference is between the decision surfaces of the two systems.



10. Why is it difficult to write classifiers for protein groups? In what sense can protein groups be different from each other?

09_Similarity_search_2a - page 22

a) Why is it difficult to work with (or to maintain) similarity groups
- **You can find groups using multiple alignment**
- Number of members from 1 to >30 thousand

• Average sequence length from 5 to several thousand
• Tight groups: all members similar with an E < 10^-4;
Loose groups, one member is similar only to a few other members
• Relation of within group to between-group similarity varies…
Similaritywithin >>Similaritybetween

b) milyen értelemben lehetnek a protein csoportok külöbözőek? nem értem ezt a kérdést, valaki magyarázza el :) szerintem az ellátott funkció alapján pl lehetnek különbözőek.-->origin, function, 3D-structure, conserved sequences...?
easy-difficult, large-small, tight-loose 09_Similarity_search_2a.pdf - page 39

11. What databases use a)annotated clusters? b) non-annotated or partly annotated clusters?

a)
COG (NCBI) - 09_Similarity_search_2a.pdf - page 20
SBASE - 09_Similarity_search_2a.pdf - page 20

b)
Uniref - 09_Similarity_search_2a.pdf - page 30

12. Describe a current solution of sequence clustering (CD-hit),explain why certain solutions were chosen.

09_Similarity_search_2a pdf 34.o

Represent sequences as an index table
• 1) For sequence i=1: retrieve all sequences that share a „sufficient number" of identical words.
• 2) Do an all-against-all for the retrieved group. Those sequences that are above threshold (say, 90 % identity),are recorded and excluded from further comparison.
• Repeat steps 1 and 2 until there are no more sequences.
This is how the clusters of the UNIPROT sequencedatabase are prepared (Adam Godzik, CD-hit programs)

b)
• Problem 1: Time All-against-all is too expensive to compute with the only suitable algorithm (string alignment with dynamic programming) - 09_Similarity_search_2a.pdf / 32
-->solution: Use an inexpensive mixed description at first, like word presence-absence vectors
• Problem 2: Memory Word-frequency vectors AND All-against-all matrices are too big (but fortunately, sparse) - 09_Similarity_search_2a.pdf / 33
-->solution2.1::Use a compressed representation for word frequency matrices a) hash table (sequence x: word1, word2,..) or b) index table (word x: seq1, seq2, ...)
-->Solution 2.2: Compute all-against-all only for the groups deemed OK by word frequency..

13. Name two major application areas of Next Generation Sequencing!

de novo genome assembly; personalized medicine; resequence known genomes from multiple individuals, ?sequencing ancient, fragmented DNA?

14. Compare three aspects of Next Generation Sequencing with the Sanger method (name the benefits/drawbacks of NGS)!

10_NGS.pdf 5.oldal

Common aspects of NGS technologies
- Steps of determining a DNA sequence (also in Sanger):
    - template preparation
    - sequencing and imaging
    - data analysis
- Assessment of base quality (also an issue in Sanger)
    - No uniform method and notation
- Relatively short (~30-400 bp) reads (compared to Sanger), efforts to increase read length
    - Putting the segments together (=assembly) is entirely a computational task
    - Short read length is an advantage for some applications (e.g. ancient DNA is fragmented and not suitable for Sanger)

15. What types of templates can be used for Next Generation Sequencing and what are they good for?

Template types:
- fragment templates: random < 1kb DNA segments
- mate-pair templates: circularized DNA fragments (e.g. ~2 kb, ends joined) cleaved subsequently (contains information on sequences further apart)

16. Name two of the biochemical principles of sequence determination by Next Generation Sequencing!

10_NGS.pdf 8. diától
cyclic reversible termination, sequencing by ligation, single-nucleotide addition

17. What is „color space”?

10_NGS.pfd / 11 alapján

A projection of sequence space. A base pair represents one color. Therefore one color can represent several “base pairs”. Base pair mint két egymás mellett lévő bázis alkotta pár.

18. What is the principle of encoding base quality in files storing sequencing output?

Each base in a sequence comes with a “quality score”: measures the probability that the base is called (=identified) incorrectly. QPHRED=-10 x log10(P), where P is the probability of the base call being wrong. Q=20 means 1 error per 100 bases, an accuracy of 99%, it is usually considered the minimum acceptable score.

19. What is the concept of next Generation Sequencing read storing by SAM format?

10_NGS.pdf 24. dia
about SAM format:
- SAM=Sequence Alignment/Map format (BAM: compressed binary version)
- Platform-independent, stores data already processed (=aligned to reference)
- Used to store short reads aligned to a reference sequence (does not contain the reference, however!)
- Contains the base sequence, mapping/alignment specifiers and quality data
- Alignment info is stored in CIGAR strings

20. Describe three aspects of targeted sequencing!

http://www.illumina.com/applications/sequencing/targeted_resequencing.ilmn
Targeted resequencing isolates genomic regions of interest in a sample library, focusing on targets and mutations. Isolating genomic regions of interest through targeted resequencing enables systematic detection of common and rare variants for high-throughput sequencing at a lower cost per sample.

21. Define metagenomics!

Metagenomics is an emerging field in which the power of genomic analysis (the analysis of all the DNA in an organism) is applied to entire communities of microbes, bypassing the need to isolate and culture individual microbial species, genetic material recovered directly from environmental samples.

22. What is ChIP-SEQ and what is it good for?

ChIP-seq combines chromatin immunoprecipitation (ChIP) with massively parallel DNA sequencing to identify the binding sites of DNA-associated proteins. It can be used to map global binding sites precisely for any protein of interest.

21-27-es kérdéskör az melyik diában van? Amelyikhez már van írva azt hol találtátok? (see top of the document.. Geri)

23. Give at least 6 functional similarities that could be the basis of interpreting a gene/protein list!

24. List experimental methods generating a gene/protein list as output!

25. The p-value of CAM pathway is 0.023, what does this p-value mean?
09_Similarity_search_BLAST.pdf 33

The p-value is the probability of observing data at least as extreme as that observed. This means, that the probability of a plant to evolve into the CAM pathway is 0.023.

26. Give a brief description of gene set enrichment analysis (GSEA)!

(GSEA) is a computational method that determines whether an a priori defined set of genes shows statistically significant, concordant differences between two biological states (e.g. phenotypes).

from http://www.broadinstitute.org/gsea/index.jsp

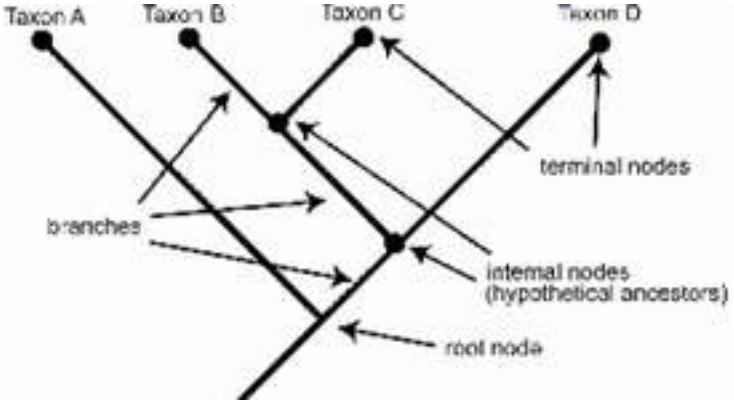
27. What is the structure of COG database?

COG is a collection of bacterial sequences, grouped by BLAST similarity.

----------a ZH1--ből ide tartozó kérdések----------

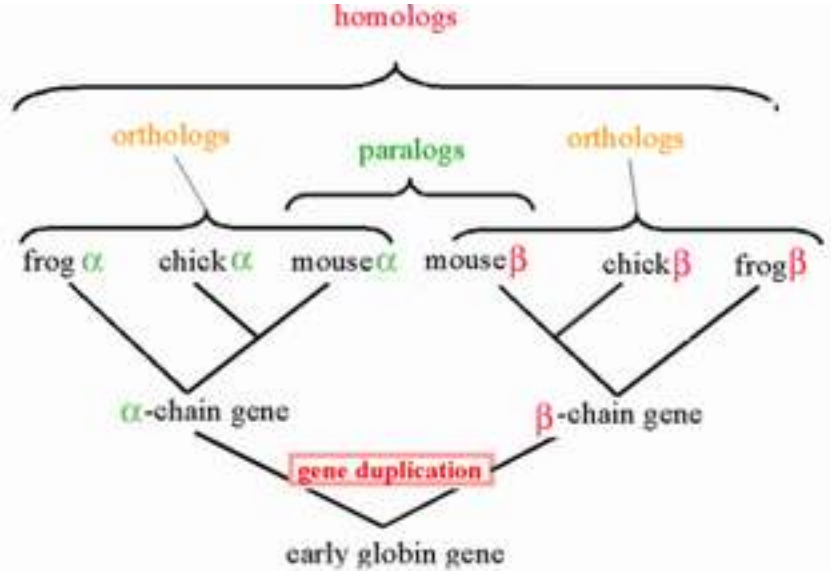
16. What is a phylogenetic tree? What is the basic assumption? (brief definition)

- az élőlények (és vírusok) leszármazását elágazásokkal ábrázoló körmentes gráf.
- levelek = OTU-k, a recens (ma élő) fajokat (vagy egyéb taxonokat) reprezentálják
- lehet gyökeres(rooted) vagy gyökértelen (unrooted)
- Az egyes élekhez pozitív számokat rendelhetünk, amelyek az evolúciós távolságot reprezentálják. Ezeket a számokat hívjuk az élek hosszának.



17. When do we say that two protein are orthologs? When are they paralogs?

- Ortológia: két gén ortológ, ha két különbözõ fajban találhatóak, és egy közös õs-génbõl származnak, mely a két faj közös õsében volt jelen. Ugyanazt a funkciót szolgálják, a két fajban.
Példa: laktát dehidrogenáz (ill. génje) az emberben és az egérben.
- Paralógia: két gén paralóg, ha ugyanabban az organizmusban találhatóak, és egy közös õs-génbõl génduplikáció és azt követõ divergens evolúció útján alakultak ki. Többnyire különbözõ, de egymással összefüggésben lévõ funkciójuk van.
Példa: a hisztidin-bioszintézis enzimei (ill. ezek génjei) emberben (nagyon hasonló szerkezetûek, de más-más reakciót katalizálnak).



18. What are the two main methods for building phylogenetic trees?

I. Distance-Based Methods: (Távolság alapú)

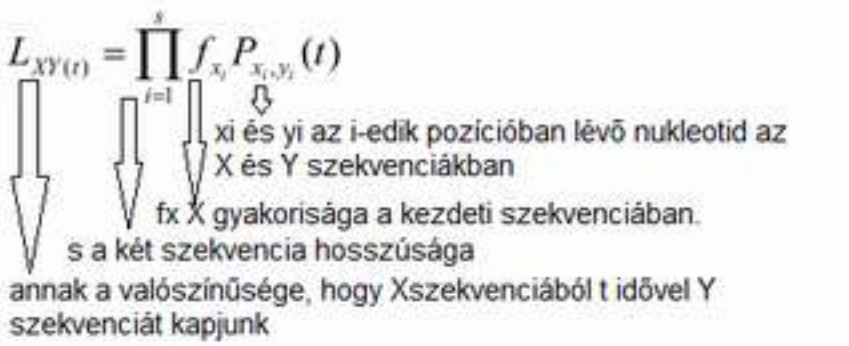
- Alapötlet: Páronkénti távolságokat számít a szekvenciák között, majd ezekkel a távolságokkal dolgozik tovább, fákat levezetve belõlük.
- Hátrány: A távolságszámításnál mindig információvesztés van, csak egy lehetséges fát ad megoldásul,
- Előny: kisebb a számítás igénye, gyors, nagy adathalmazokkal is elbír
- Távolságmátrixból -->fa módszerek:
  Klaszterezés (clustering):
  agglomerative: elejében minden elem külön csoportba tartozik és lépésenként összevonjuk a legközelebbieket
  divisive: elejében minden elem egy csoportba tartozik és ezt lépésenként bontjuk részcsoportokra
  - Least squares (LS) method
  - Minimum evolution (ME) method: A legrövidebb olyan fát találja meg, amely összeegyeztethetõ a szekvenciák közötti távolságokkal.
  - UPGMA:gyökeres fát ad, agglomeratív, csoportátlag alapú eljárás
  - Neighbor Joining (NJ): Egy csillag alakú fából kiindulva a legközelebbi szomszédokat összekapcsolja, helyettesíti õket az átlagukkal, majd ezt ismételgeti a teljes fa kialakulásáig.

II. Character-Based Methods (Karakter(= pozíció az összerendezésben) alapú)

- Alapötlet: Megtartják a taxon eredeti szekvenciáját, ezzel rekonstruálhatóvá téve a nóduszok, vagyis az ősi fajok lehetséges szekvenciáját.
  A törzsfa elkészítéséhez nemcsak a szekvenciák közötti távolságra, hanem a szekvenciákban levő egyéb információkra is van szükség.
  Olyan fákat származtat le, amelyek mindegyik pozícióra optimalizálják az adatmintázatok eloszlását. (a legkisebb számú helyettesítést igényelnek)
- Módszerek a faépítésre - példák:
  - Maximum Parsimony (MP): "legnagyobb takarékosság" módszere: Olyan fát épít, ami a lehetõ legkevesebb evolúciós eseménnyel (mutáció) magyarázza meg a meglévõ szekvenciák létrejöttét. Jellegek: Ezen tulajdonságok segítségével különböztetjük meg az OTU-kat. Lehetnek folytonosak vagy diszkrétek, az OTU-kon vizsgált jellegek vagy karakterek állapotait egy taxon-tulajdonság mátrixba gyűjtjük.
    Számos azonos pontszámú fát szolgáltat, ezek közös részét vehetjük mint megbízhatót. Arra kell törekednünk, hogy a fa a legtöbb homológiát (=A tulajdonságok közös eredetén alapuló valódi hasonlóság) és a legkevesebb homopláziát(=Ha egy jellegállapot két faj között hasonló, de valójában csak a konvergens evolúció következménye) tartalmazza.
    Nagy távolságú szekvenciák esetében hátránya, hogy azonos bázis esetén azt tételezi fel, hogy nem történt mutáció, holott valószínûbb a visszacserélõdés.
  - Maximum Likelihood (ML): "legnagyobb valószínûség" módszere: Komplikált módszer. Minden pozícióra kiszámítja, hogy adott fa és helyettesítési modell mellett mi a valószínûsége annak, hogy a megfigyelt variációs mintázat jöjjön

létre az adott pozícióban. Az egyes pozíciókra kapott valószínûségek összeszorzásával adódik a teljes fa valószínûsége. Ezt sok fára a legjobbat kiválasztja. Ezt többféle helyettesítési modell mellett is elvégezhetjük, ezek közül is kiválasztva a legjobbat.

A maximum likelihood eljáráson alapuló algoritmusok tehát logL maximalizálására törekednek.



What is the probability of seeing the observed data (D) given a model/theory (T)? Pr(D|T)

Igen számításigényes, de ez a legmegbízhatóbb.

- Bayesian inference

  A módszer alapjául az apriori (előzetes) és a posterior (utólagos) valószínűségek szolgálnak.

  Az MCMC fakeresés lényege:
  A Markov lánc valószínűségi változók olyan sorozata, mely minden egyes tagjának a valószínűsége csak a sorozat előző tagjától függ.
  Keresi azokat a fákat, melyekre igaz, hogy az adott adatok (szekvenciák) mellett maximális a likelihoodja. A kapott likelihoodokat a Bayes tétel segítségével valódi valószínűségekké alakítja, hogy a valószínűségek összege egy legyen, és így tovább lehessen számolni velük. Nem egy legvalószínűbb fát keres, hanem sok nagyon valószínű fának a konszenzusa lesz a végeredmény.

19. Describe the principle of the UPGMA method?

UPGMA (Unweighted pair-group method with arithmetic mean)

i. Adott egy távolságmátrix. Tfh. a legkisebb távolság A és B között van, ezért OTU(A,B) klaszterbe összevonom őket.

ii. Újra számolom a távolságmátrixot

- A és B helyett már OTU(A,B)-vel fogok számolni
- d(OTU(A,B),OTU(C))=[d(OTU(A),OTU(C))+d(OTU(B),OTU(C))]/2
  Az összevont OTU és a sima OTU közötti távolság egyenlő a sima OTU és az összevont OTU-ban lévő sima OTU-k távolságának átlagával.

iii. Akkor van vége az algoritmusnak, ha az összes elem belekerült a fába, mindegyik OTU lett

(OTU: Operational Taxonomic Unit, „kezelendő taxonómiai egység”. Általában fajokat vagy magasabb taxonómiai egységeket jelölnek)

20. Here is the basic equation of the Karlin-Altschul statistics: E=K*m*n*e^(-lambda*S). Explain the meaning of the formula and list the meaning of its components! Which statistical distribution is it based on?

The Karlin-Altschul statistics is based on Extreme Value distribution. The expected number of local alignments of “high scoring pairs” (HSPs), with score at least S is

E=K*m*n*e^(-lambda*S)

m is the query length, n is the entry length (taken as the whole database), K and l are constants. mn is called the search space. It is based on χ^2**-distribution**