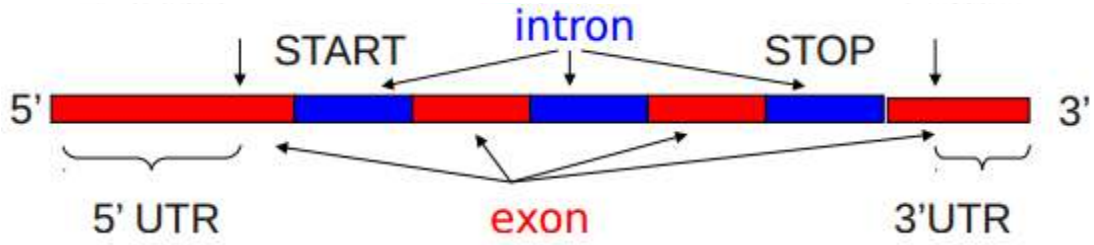
Introduction to bioinformatics, 2011.1.ZH

1. Name 4 major application areas of bioinformatics!

Description, management, interpretation (analysis, data-mining) , modelling (simulation)

2. Name the parts of an eukaryotic gene and describe BRIEFLY their function!

Diában csak ezt találtam:



Ez persze nem az igazi, az 5' UTR- ben vannak promóterek, ill a TATA box is

3. Name 4 core data types of bioinformatics!

Sequences

3-D

Networks

Text.

(Each (except text) has an underlying model, a core description, plus various simplified and/ or extended (annotated) descriptions.)

4. Define Hamming distance and Levenshtein's edit distance!

The Hamming distance is the number of exchanges necessary to turn one string of bits or characters into another one. It is assumed that the two strings are of identical length and that no alignment is necessary.

Levenshtein distance can be defined as a sum of costs assigned to matches, replacements and gaps. The two strings do not need to be of the same length.

How would you calculate the two distance values for the following sequences?

V = ATTAATAT ;

W = TTTATATA

Hamming distance: 5

~~//Levenshtein distance: 5 (de javítsatok ki, ha nem!)~~

http://www.functions-online.com/levenshtein.html (vigyázz, rossz a tool, a sor végi szóközt beleszámolja!!)

~~E szerint a levenshtein csak 2, de logikus is, mivel az első karaktert kicseréled, ill W-ben egy gap-et teszel az első A (4. karakter) elé, vagy után!~~

Szerintem meg **3 lesz a levenshtein**, mert az elsőt cseréled, van egy insert vagy egy gap és a végén még van még egy karakter amit el kell tüntetni.

A program szerint is három, nem kettő...

5. What are the two descriptor types in the annotation of a protein or DNA sequence data in a sequence database? Give an example of each!

Global descriptors e.g. function

Local (positional) descriptors e.g. binding sites, domains

6. List three important properties of the SwissProt database!

Summary of the current knowledge on a given protein

Proteins are selected for manual annotation based on curation priorities

UniProtKB/Swiss-Prot gathers data from multiple sources

7. Which of the databases below is primary?

(GenBank, CATH, PDB, SwissProt, SCOP, SBASE, RefSeq, Uniref50, Superfamily, InterPro)

GenBank, SwissProt, PDB

8. Gives three possible reasons why you might not find the (correct) data you need in a specific database!

after all they are maintained by human beings

there are experimental errors (even kinds that are hard to detect)

the ultimate verification of an experiment is its reproduction - clearly not feasible for all database depositions

there can even be intentional depositions of false data (cheating, H.M. Krishna Murthy)

9. Describe the main idea of dynamic programming in a few sentences!

Dynamic programming is a programming paradigm that stores sub-results in memory, uses them for later steps instead of recalculating the values and thus saves runtime.

10. What kind of special handling of gaps was introduced in sequence alignment and why?

score for a gap of length x is:

$-(\rho + \sigma x)$

where $\rho > 0$ is the penalty for introducing a gap

(gap opening penalty) and

$\sigma > 0$ is the penalty for extending a gap

(gap extension penalty)

ρ will be large relative to σ because you do not want to

add too much of a penalty for extending the gap.

//Egy nagyobb géndarab nagyságrendileg kb azonos eséllyel esik ki, mint amilyen egy pontmutáció valószínűsége

11. What is the difference between global and local alignment?

Fill in the matrix with max value of 4 possible values:

0 (ha a gap penaltyvel negatív értéket kapnánk)

Vertical move: Score + gap penalty

Horizontal move: Score + gap penalty

Diagonal move: Score + match/mismatch score

Minden elem >=0, így pozitív/0 értéket kapunk

-erre szerintem ilyesmi a válasz:

Global alignment: Align two sequences from “head to toe”. Local alignment: Locate region(s) with high degree of similarity in two sequences.

12. Prepare the global alignment of the words 'APPLE and 'HAPPY' using the following matrix and a penalty -1 for indels!

Ha

GAP:-1, match:1, miss: -1

Szerintem:

	-	A	P	P	L	E
-	0	-1	-2	-3	-4	-5
H	-1	-1	-2	-3	-4	-5
A	-2	0	-1	-2	-3	-4
P	-3	-1	1	0	-1	-2
P	-4	-2	0	2	1	0
Y	-5	-3	-1	1	1	0

\- A P P L E

H A P P Y - De a mátrix alapján több megoldás is lehet

13. Generally why is not practical to extend the dynamic programming solution of pair-wise alignment for multi-alignment of k sequences? What other algorithmic option can be used instead?

k sequences need a k-dimensional matrix

For k sequences, build a k-dimensional Manhattan, with run time (2^k-1)*(n^k)

O(2^k*n^k)

impractical due to exponential running time

The most widely used progressive alignment algorithm is currently CLUSTAL W.

14. How are amino acid substitution matrices constructed? What is the difference between PAM and BLOSUM matrices?

(5.eloadas 13-15 dia)

It is a 20x20 symmetrical matrix that can be constructed from pairwise alignments of related sequences

“Related” means either

a) evolutionary relatedness described by an “approved” evolutionary tree (Dayhoff’s **PAM** matrices)

Derived from the BLOCKS database

b) any sequence similarity as described in the PROSITE database (Hennikoff’s **BLOSUM** matrices)

Derived from manual alignments of closely related proteins

(A másik jegyzet szerint Blosum = BLOCKS substitution matrix, szóval óvatosan. A Pam néhol Point, néhol Percent Accepted Mutation.)

Similarity := low PAM, high BLOSUM #

15. What is the main difference between nucleotide and amino acid substitution matrices? Write a few words about a nucleotide substitution matrix!

?az aminoban figyelembe veszik az egyes aminosavak cseréjének esélyét?

4x4 vs 20x20? Hát ez biztos XD - még ez sem :) az (4+1)x(4+1) és (20+1)x(20+1) lesz szerintem

To generalize scoring, consider a (4+1) x(4+1) scoring matrix δ. The addition of 1 is to include the score for comparison of a gap character “-”.

This will simplify the algorithm as follows:

$$s_{i,j} = \max \begin{cases} s_{i-1,j-1} + \delta(v_i, w_j) \\ s_{i-1,j} + \delta(v_i, -) \\ s_{i,j-1} + \delta(-, w_j) \end{cases}$$

BTW, PAM and BLOSUM are similarly scoring matrices (size of (20+1)x(20+1)) for amino acid sequence alignments

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

● A tree will reflect phylogeny only if it is based on evolutionarily meaningful characters

● Known organisms (or proteins) are always at the tips of the branches

● Common ancestors at nodes are always hypothethical

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).

forrás:http://www.enzim.hu/~szia/bioinformatika/ea01/ea01.htm

Orthologs and paralogs are both homologs. O:speciation P:internal duplication

In other words each of them has common ancestor (homology), but paralogs have similar function, orthologs don't.

http://www.bio.davidson.edu/Courses/Molbio/MolStudents/spring2010/Rydberg/Orthologs.html

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

Distance-Based Methods

Character-Based Methods

google: http://guava.physics.uiuc.edu/~nigel/courses/598BIO/498BIOonline-essays/hw2/files/hw2_li.pdf

a. Phenetics (distance based)

The study of relationships between a group of organisms on the basis of the degree of similarity between them

b. Cladistics(character-based)

The study of the pathways of evolution

19. Describe the principle of the UPGMA method?

UPGMA (Unweighted pair-group method with arythmetic mean)

e. UPGMA (Unweighted pair-group method with arythmetic mean)

i. Cluster pair with smallest distance

ii. Recalculate distance matrix

iii. Calculate new distance using composite OTU(A,B):

iv. Distance between a simple OTU and a composite OTU is the average of the distances between the simple OTU and the constituent simple OTUs of the composite OTU

v. Warning: UPGMA is simple but often makes mistakes

//Abból az anyagból amit Klimaj Zoli feltett

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.

21. Bonus question: here

and here