lab5_questions_and_maj_answers

Course: Genomics & Bioinformatics

Lab: 5 - Alignments of mitochondrial sequence

PART 1

Author: Mohammed Al-Jaff

Part 1: Pairwise Alignment

Q1.1:* Which sequence format are the two sequences listed in?*

Ans: Format looks like a traditional fasta format with multiple sequences and with a amino acid 'alphabet'.

Q1.2: Report the following values / observations from the alignment.

- Alignment score
- Alignment length
- % and fraction Identity (The value reported for "Identity" includes perfect matches only)
- % and fraction Similarity (The value reported for "Similarity" includes perfect matches + "close" mismatches)

Ans: Global alignment pairwise results:

```
- Alignment score : 860.5
- Alignment length : 361 a.a. long (?)
- % and fraction Identity: 48.8%
- % and fraction Similarity: 25.5%
```

Q1.3: Report the same values as above (Alignment score etc). Consider the alignments produced by the two different approaches: do YOU think one of them is more biologically relevant than the other, or do both contribute valuable information?

Ans: Local alignment output:

```
- Alignment score : 916.0
- Alignment length : 269
- % and fraction Identity: 65.4%
- % and fraction Similarity: 79.6%
```

Given that this could be a case where a whole initial chunk of the protein was

By visual inspection, it almost looks like booth alignments actually are really the same but look on first impressions different because we see the begining subsequence chunk of 'above' protein not being aligned [but shown], while in the local alignment with Water, that part of the first protein is simply discarded and what is shown is the best alignment of two subsections between the two proteins [makes sense, given a *local* alignment].

Q1.4: Let's go a bit deeper into why the two sequences differ in the N-terminal part: Look up both entries in UniProt (http://www.uniprot.org) and try to locate information regarding the following questions.

i. How were the amino acid sequences of the two proteins determined? (Hint: look at the titles of the papers, and the Cited for fields, listed in the Reference sections).

ii. Subcellular localization: Where in (or outside) the cell do the enzymes function?

iii. The feature table contains details about the regions/domains of the protein - try to do a comparison to spot the differences between the two UniProt entries (Hint: focus on the "PTM / processing" section).

Ans

i.

The earliest publication for P29600 - SUBS_BACLE indicates that both the sequence and 3d structure of the protein was determined by crystallography. For P41363 - ELYA_BACHD, the sequence of at least the first(N-terminal) 20 aa sequence was determined biochemically with an *"Applied Biosystems Protein/Peptide Sequencer, model 477A"*. ["Molecular cloning, nucleotide sequence and expression of the structural gene for a thermostable alkaline protease from Bacillus sp. no. AH-101." by Takami et al. Appl. Microbiol. Biotechnol. 38:101-108(1992)]

ii.

P29600 - SUBS_BACLE: "Subtilisin is an extracellular alkaline serine protease, it catalyzes the hydrolysis of proteins and peptide amides."[uniprot] Ie produced by the organism and secreted out during sporulation.

P41363 - ELYA_BACHD - Also extracellular protein that is secreted according to its GO classification.

iii.

P41363 - ELYA_BACHD: Consisistes of 3 parts that undergo processing (perhaps during its way tora) as shon in the below taken from uniprot:

```
1. Signal peptidei:     1 – 24 (24aa long)
2. Propeptide: 25 – 93    1 (69aa long )
3. Thermostable alkaline protease: 94 – 361   (268aa long)
```

P29600 SUBS_BACLE: This particular protein entry consists of only one component of length 269aa. Observe that that this corresponds to the more or less equivalent sequence to the 3rd protease component of the above P41363 - ELYA_BACHD*.[minus 1 aa?].

Given our alignments [both global and local] we hypothesises that the alignment is legit and that the difference in sequence appearance between the two sequences/proteins are due to the larger one being 'whole' pre-porccesed version of the *P29600 SUBS_BACLE*. When performing christalogrphy *P29600 SUBS_BACLE*, may have been purified in its extracellular processed/cleaved form.

Q1.5: Based on what you've learned about the P41363 protein from the alignment to Savinase and from the data on the Uniprot site: do you think this could be used as an enzyme in washing powder? (Why? / why not?).

Ans I think my default position is 'No'ish' since there might be an additional post translation processing step stage which might or might not be practical industrially to deal with and because P41363 seems to be a 'niche' protease that targets serine residue while the other protein is mentioned as a "broad spectrum protease" with a larger and more flexible cleave ability, meaning that more 'dirt' can be degraded.

Section 1. part 2 - alignment of two "dissimilar" sequences

Q1.6: Compare Savinase to the human peptidase by global alignment (Needle) — remember again to set End Gap Penalty to "true" — and report the following: Alignment score, Alignment length, Identity, Similarity, How large a part of the alignment is gaps?

Ans

```
    - Alignment score: -244
    - Alignment length: 1225
    - Identity: 8.8%
    - Similarity: 12.3%
    - How large a part of the alignment is gaps?: 79% (!)
```

Q1.7: Repeat the alignment with End Gap Penalty set to "false" and report the same results as above.

Ans

```
  - Alignment score:
  - Alignment length: 1289
  - Identity: 5.7%
  - Similarity: 10.2%
  - How large a part of the alignment is gaps?: 82.2% (!!)
```

Q1.8: Repeat the alignment again — this time using the local alignment algorithm (Water) — and report the same results as above.

Ans

```
  - Alignment score: 173.0
  - Alignment length: 296 a.a
  - Identity: 24.0%
  - Similarity: 43.6%
  - How large a part of the alignment is gaps?: 24.7%
```

Q1.9: Do you think local or global alignment is best for finding similar parts of distantly related proteins? Why? Hint: Distantly related proteins typically share a core, that relates to the function of the protein..

Ans As the hint indicates, distant proteins share some set of subsequences/domains with specific functionality, while those "in-between" regions act as linkers and might be more tolerant of mutations. I think that if we are dealing with single domain proteins then local alignment will aim at finding where within a given region one domain 'klicks' and aligns best inside the another sequence. Contrast this with a global alignment approach where depending on the choice of params, the algorithms tries to stretch out one sequences to align with another one while taking into account all mismatches and gaps.

However, when comparing two distant proteins where each have multiple domains that also are "distant" than a global approach with a false gap ending penalty might actually be a good compromise since the algorithm will aim to align all the multiple domains to their corresponding counterpart in the other sequence. The idea here is that the alignment where all/most domains aligning together will have the highest score despite penalty added by the unconserved linker sequences being mismatches/gaps.

Q1.10: How do the local alignments look? (What are the ranges of Alignment score, Alignment length, Identity, Similarity, and gap percentage)?

Ans Short answer; Hapharzard local alignments. Looks like given a short enough length, you can in principle fit any small section to another section with a not too bad score. Problem here is then the actual local alignment length as a percentage of the whole, because shorter sequences that happen to randomly align ok'ish can bias the identity and similarity percentage for the local alignment.

```
    In terms of Ranges for 3 shuffled local alignments

    - Alignment score: [40, 60]
    - Alignment length: [100, 275]
    - Identity: [21%, 25%]
    - Similarity: [33%, 45%]
    - How large a part of the alignment is gaps?: [26%, 39%]
```

Q1.11: Comparing the Savinase/shuffled alignment to the previous Savinase/Human Peptidase alignment - how will you judge the alignment with human peptidase now? (More/Less confidence in relation between the sequences?).

Ans: The above shuffle experiment has really made me feel less confident of the initial original results due to those results on second case being worryingly close with the 'randomised'/reshuffle ranges.

Section 1. Part 3 - about parameters

"alignments are dependent on parameters"

Q1.12: *What are the alignment results (Length, score, gaps, identity, similarity)? How do alignment length and % identity depend on the BLOSUM number (compare also to your answer to question 8)?*

Ans

Matrix: BLOSUM30 substitution matrix local alignment

```
Length: 326
Identity:      76/326 (23.3%)
Similarity:   149/326 (45.7%)
Gaps:          88/326 (27.0%)
```

```
Score: 342.5
```

Matrix: BLOSUM90 substitution matrix local alignment

```
Length: 279
Identity:      73/279 (26.2%)
Similarity:   107/279 (38.4%)
Gaps:          91/279 (32.6%)
Score: 147.5
```

Obs1: On first glance, the numbers don't tell me much until I read up on how one constructs usch BLOSUM matrices, the sequence data used to construct them and what the numbers 62/30/90 mean.

Basically, the entries is a "score"/"cost"/"value" that the local alignment also uses to assign a score for each base position between two sequences when calculating the overall score of the alignments and for assigning the values in the dynamic programming matrix.

The important and practical thing here is to realise that the numbers 62/90/30 indicate the how the threshold similarity value for the sequences used to construct the matrix. For example, BLOSUM62 is constructed by using a bunch of sequences 'known'/'seen' to have less than 62% similarity when being manually aligned. Because of this, the idea here is to use a substitution matrix with a low bloom number when you expect to align very distant sequences. For very distantly related sequences you would think/expect that similarity/identity will be lower and alignment length low too in the case of local alignment.

Q1.13: *How do the quality parameters look this time (Length, score, gaps, identity, similarity)? Is this alignment biologically meaningful at all?*

Ans local alignment with minimal gap penalty and gap extension penalty:

```
Gap_penalty: 1.0
Extend_penalty: 0.0
#
Length: 1254
Identity:     192/1254 (15.3%)
Similarity:   228/1254 (18.2%)
Gaps:        1010/1254 (80.5%)
Score: 896.576
```

None of the scores/values look good at all because they seem to be evan worse than in the random shuffle alignment. As expected, there is a huge number of gaps since there is no cost associated with introducing and extending them. What the alignment boils down to is that it seem to desperately try to align a residue in one of the sequences to the closest identical residue in the other and adding as much gaps it needs to avoid any residue mismatches.