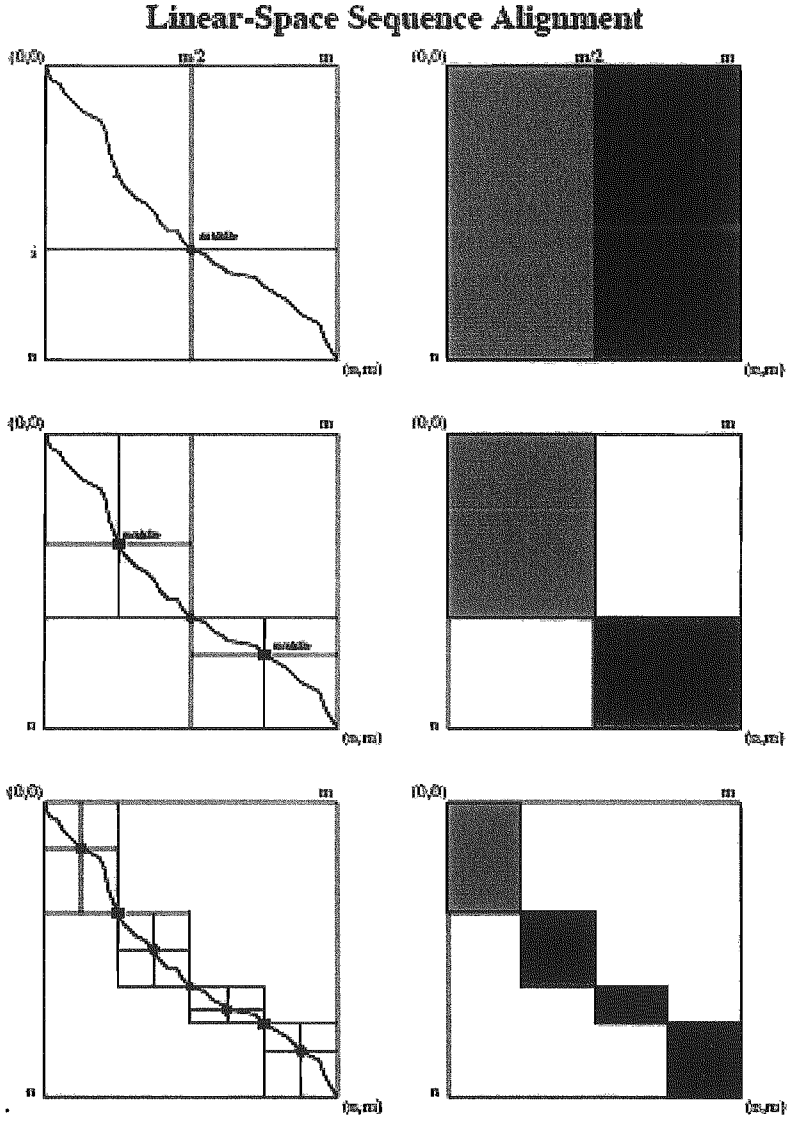
Question 1

(a) For the Smith-Waterman Space the complexity for computing alignment path for sequences of length n and m is O(nm). We need to keep all backtracking references in memory to reconstruct the path (backtracking); while the Space complexity of computing just the score itself is O(n), n>m. We only need the previous column to calculate the current column, and we can then throw away that previous column once we're done using it. The figure show how to find the middle point which wil lead to store only O(n) values.



The algorithm can be briefly described as follows:

Path $(source, sink)$
if $source$ and $sink$ are in consecutive columns
    output the longest path from the $source$ to the $sink$
else
    $middle \leftarrow$ middle vertex between $source$ and $sink$
    Path $(source, middle)$
    Path $(middle, sink)$

(b) BLAST (Basic Local Alignment Search Tool) is created to do similarity searches of a query against a database. The first preparatory step consists of constructing a hash table of all seeds occurring in the query sequence. A seed of weight k is a word consisting of k contiguous nucleotides (k-word),. The search needs to be precise, i.e. should report all, or at least a vast majority of interesting similarities that could be relevant in the underlying biological study. This requirement for the alignment method, is called sensitivity and, counterweights the speed requirement, usually directly related to the selectivity (called also specificity) of the method. The fundamental unit of BLAST output is the High-scoring Segment Pair (HSP). The HSP consists of two sequence fragments of arbitrary but equal length whose alignment score meets or exceeds a threshold score. The standard BLAST algorithm parameters are word length w, word score threshold T and segment score threshold S. The approach to similarity searching used by the BLAST program is first to look for similar segments (HSPs) between the query sequence and a database sequence. This begins with identifying short words of length w in the query sequence that either match or satisfy some positive-valued threshold T when aligned with a word of same length in a database sequence. This is done by building an automaton of all the neighbors of the words. These hits act as seeds for initiating searches to find longer HSPs containing them. The word hits are extended in both directions along each sequence for as far as the cumulative alignment score can be increased. Extension of the word hits are halted when: the cumulative alignment score falls off by the quantity S from its maximum achieved value or the end of either sequence is reached. PSI-BLAST (Position Specific Iterated ? BLAST) is the state of the art Blast software. Iterative Procedure: Performs BLAST on a database and Uses significant alignments to construct a position specific score matrix. This matrix is used in the next round of database searching until no new significant alignments are found. PatternHunter was the first method that used carefully designed spaced seeds to improve the sensitivity of DNA local alignment. A spaced seed is formed by two words, one from each input sequence, that match at positions specified by a fixed pattern – a word over

symbols # and _ interpreted as a match and a don't care symbol respectively. For example, pattern ##_# specifies that the first, second and fourth positions must match and the third one may contain a mismatch. Spaced seeds have been shown to improve the efficiency of lossless filtration for approximate pattern matching, namely for the problem of detecting all matches of a string of length m with q possible substitution errors (an (m, q)-problem). Other software use some specific spaced seeds and random spaced seeds

Question 2

(a) Non-biological "background" signal, like non-specific hybridization and technical noise, might contribute to the measured intensities ("foreground"). Side-by-side boxplots of the distributions of background and foreground signals for both channels (for the case of two-colour microarrays) in all arrays can help to assess the overall background effect. Plots of spatial distribution for background and foreground help to detect spatial artefacts. If background correction is found to be needed, looking at the low intensities side on MA-plots (average log-intensities vs. log-ratios) and density plots (smoothed histograms) of intensities before and after correction can help to monitor its effectiveness and the amount of noise/bias introduced by it.

For two channel microarrays, there is usually some dye bias and the need for within-array normalisation of intensities, so the two channels can be compared. Median normalisation (i.e making the median intensities of both channels equal, which is equivalent to setting the median M to zero) can be applied if one assumes that the changes are overall roughly symmetric. If there is an intensity-dependent dye bias (which can be monitored with the MA-plot), "loess" normalisation (i.e. equalling the loess line of M-values across intensities to zero) is more appropriate and assumes that the changes are roughly symmetric at all intensities. If there is information about the printing of the array and an effect associated with the print-tips is detected, "loess" normalisation can be performed for each print-tip group of probes. Quantile normalisation is usually used between single-channel arrays and relies on a stronger assumption, enforcing the chips to have identical intensity distribution. Boxplots and density plots of intensities can be used to monitor the effectiveness of the normalisation procedure.

References:

Lecture 2 notes;

Smyth GK, Yang YH, Speed T. "Statistical issues in cDNA microarray data analysis". Methods Mol Biol 2003;224:111-36. [PMID: 12710670];

Smyth GK, Speed T. "Normalization of cDNA microarray data". Methods 2003 Dec;31(4):265-73. [PMID: 14597310].

(b) A design matrix must be created to reflect the expected log-ratio for each slide. These log-ratios must be a linear combination of independent contrasts of interest (parameters). Let X be the design matrix and σ the standard deviation between slides for a particular gene. The standard error of the ith parameter estimate is then given by

$\sigma\sqrt{c_i}$, where c_i is the ith diagonal element of the matrix $(X^TX)^{-1}$. These calculations assume independence of replicates, which does not happen in reality. There is always some correlation between the expression levels of different sample types. Moreover the effective replication for each sample type depends on the correlation between replicates. The more independent the replicates, the higher the effective replication. Having highly correlated replicate arrays is equivalent to having fewer arrays on the estimation of the parameters and their variances.

References:

Lecture 2 notes;

Glonek GF, Solomon PJ. "Factorial and time course designs for cDNA microarray experiments". Biostatistics 2004 Jan;5(1):89-111. [PMID: 14744830].

(c) Relative gene expression measurements (observed M-values) canâĂŹt be assumed to result from a normal distribution. M-values actually have a very heavy tailed distribution. Moreover, the sample size for each gene corresponds to the amount of replication for each type of array, which is usually very small. ItâĂŹs hard to judge significance on small sample sizes, as the assumptions for any statistical test become weak. Moreover, microarray experiments typically have thousands of genes, whose M-values have different variances and are correlated in an unknown way, and so there is a high level of multiple testing.

The B-statistic moderates the standard t-statistic by incorporating information about the variability of all the other genes to smooth the standard error for each individual gene. The higher the number of genes and/or the lower the number of replicates, the stronger the "shrinkage" of the standard error. The B-statistic benefits from the best of the t-statistic, which is to prevent the average M to be driven by outliers (specially for a small sample size) by incorporating information about its variability. It also avoids the main problem of the t-statistic, which is to be driven by tiny variances that are likely to appear randomly when many tests are performed, by empirically smoothing the standard error. The B-statistic ranks the genes according to the evidence (log odds) for differential expression.

References:

Lecture 4 notes;

Smyth GK. "Linear models and empirical Bayes methods for assessing differential expression in microarray experiments". Stat Appl Genet Mol Biol 2004;3:Article3. [PMID: 16646809].

Question 3

(a)(i)

$$\frac{dP(k,t)}{dt} = f(k-1)P(k-1,t) + g(k+1)P(k+1,t) - [f(k)+g(k)]P(k,t)$$

(ii)

$$\begin{aligned}\frac{d\langle x\rangle}{dt} &= \sum_k k\frac{dP(k,t)}{dt} \\ &= \sum_k (k+1)f(k)P(k,t) + \sum_k (k-1)g(k)P(k,t) - \sum_k k(f(k)+g(k))P(k,t) \\ &= \sum_k f(k)P(k,t) - \sum_k g(k)P(k,t) \\ &= \langle f(x)\rangle - \langle g(x)\rangle\end{aligned}$$

At equilibrium $\frac{d\langle x\rangle}{dt} = 0$ hence $\langle f(x)\rangle = \langle g(x)\rangle$.

(iv) At equilibrium $\frac{d\langle x^2\rangle}{dt} = 0$ and $\langle f(x)\rangle = \langle g(x)\rangle$. Note that using the linearisation we also have $f(\langle x\rangle) = g(\langle x\rangle)$. The ODE for $\langle x^2\rangle$ gives

$$\begin{aligned}&2\langle x[f(\langle x\rangle) - g(\langle x\rangle) + (x-\langle x\rangle)(f'(\langle x\rangle) - g'(\langle x\rangle))]\rangle + 2f(\langle x\rangle) = 0 \\ &\Rightarrow (x-\langle x\rangle)^2(f'(\langle x\rangle) - g'(\langle x\rangle)) + f(\langle x\rangle) = 0 \\ &\Rightarrow \sigma_x^2 = (x-\langle x\rangle)^2 = \frac{f(\langle x\rangle)}{g'(\langle x\rangle) - f'(\langle x\rangle)}\end{aligned}$$

Approximation is valid if fluctuations are small (e.g. in large molecule numbers).

(b)

$$f(y) = \frac{K}{K+y^h}, \quad g(y) = \beta y$$

$$\frac{df}{dy} = \frac{-Khy^{h-1}}{(K+y^h)^2} = -h\frac{y^{h-1}}{K}f^2(y)$$

From (a) at steady state $f(\langle y\rangle) = g(\langle y\rangle) = \beta\langle y\rangle$. So

$$\sigma_y^2 = \frac{\beta\langle y\rangle}{\beta + h\frac{\langle y\rangle^{h-1}}{K}(\beta\langle y\rangle)^2}$$

So increasing h decreases σ_y^2 for $\langle y\rangle \geq 1$ and not necessarily true otherwise, in which case the approximation is very likely not be valid anyway.