LAST找可变长度种子，扩展用X-drop算法初始命中gapped比对

LAST finds variable length matches instead of fixed-length seeds depending on the number of occurrences of each match. It then extends initial hits with gapped alignments using X-drop algorithm.

BLASR(12)使用后缀数组或bwt - fm数据结构来索引目标序列，然后在每个reads位置找到比最长公共前缀(LCP)更短的精确匹配。BLASR根据种子的稀罕性选择顶级的得分链来执行稀疏的动态规划(SDP)，然后利用高分的SDP比对路径限制搜索区域进行详细的带状动态规划。

BWA-MEM设计为长序列匹配，支持单端和双端reads，用Z-dropoff 带状动态规划扩展最佳位置

COSINE计算长为w的窗口中短的k-mers（3-4）出现的数量

在一个长度为 l 的序列中，一个窗口出现在*i* 位置



where

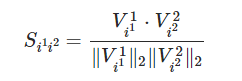






q是四个字母表的指数排列情况，相当于4^k 个取值，是具体的每个k-mer取值，如k=2时，可取AA，AC等，共有q种取值，然后nj表示每个k-mer的计数。

那么两个序列局部相似性是

这里序列1代表查询序列，序列2表示参考序列，

两条长度分别为 l1 和l2 的序列之间总相似度得分，是在偏移量m下，求取S的平均值，假设位置是等距选择，距离为d

在实际序列中，由于插入删除，两条序列偏移不一定都是m，可能偏移量是各种不同的m~，那么k-mers的分布作为每个窗口相似性度量，这里用余弦相似度好处是能够用傅里叶变换简单实现计算S

前期模拟了大肠杆菌基因组中1000个位置，然后将序列中随机产生15%-35%的突变版本，图4结果表明提取的序列与其突变的版本具有比从基因组中不相关的位置更高的相似性，但是对于更高的误差率为35％，差异变得不太可区分。

我们观测到S. coelicolor 高GC含量的基因组增加了非相关位置的余弦相似度，会对两个种分布的分离有不利影响，尤其在错误率高达35%

we also observe that the high GC content of S. coelicolor genome, increases the overall cosine similarity of unrelated locations, which adversely affects the separation between the two distributions, especially with higher error rate of 35%.

为增强区分性，COSINE用一系列沿着序列的间隔窗口计算相似度作为两个序列中相关联窗口间k-mer计数向量的相似度均值

To enhance the discriminatory power, COSINE uses series of spaced windows along the two sequences and computes the similarity score as the average over cosine similarities of *k*-mer count vectors between associated windows from the two sequences.

COSINE精度主要依赖k-mer的尺寸，窗口大小w和窗口移动距离d

COSINE accuracy mainly depends on the choice of k-mer size (*k*), window size (*w*) and windows shift (*d*).

小的d能增加准确率，但是当d达到窗口大小一半时也不会显著影响比对性能，用大的窗口能增加在高插入删除比对敏感性

For three *E. coli, S. coelicolor* and human genomes, [Supplementary Figures S3–S8](https://oup.silverchair-cdn.com/oup/backfile/Content_public/Journal/nar/45/14/10.1093_nar_gkx511/1/gkx511_Supp.zip?Expires=1505960662&Signature=FXKgERyM~pvkYhoKspAINxty~b5YgyzUWIwtNkb9A1YRmwsZvQQlePICRYbs9qxqB5WLbuP3~g-wrDW8ZHhzp3~kLwwFRGJIHkuNLocbGatLe3tgB~UZmTQJX-AhIcgIuiiPtN145vJJpjkYDkgNYER5NKnZ98wfsAwfIEuLV9esXo90X~19d8At4WEVRjPicmo-fyJ9bvRGFcACpdiXJjn6LFuv7cMMZUiTVXutDKatGOHqhViL08hJXtgf0LG-UlTg0cIDXiamOmgPBI3r6BQ01iFoKIAV7go1dbB7373kfbs~VADds7-nZgndLG2NQ~65sikUdeCZJhVMp2VM6g__&Key-Pair-Id=APKAIUCZBIA4LVPAVW3Q) represent the above-mentioned results with different window sizes of *w* = 50, 100, 500 and 1000 and two different simulated error patterns (sub:ins:del proportions) of 34:33:33 with equal insertion and deletion rates and 10:62:28 with about two times insertion rate compared to deletion, which is typical for PacBio reads

这里不是直接