符号说明：

L：单个DNA序列的长度，即100；

k：子列长度，满足1<=k<=100；

N：DNA序列总个数，即1000000；

Ln：在所有序列中出现过的k-mer子列种类数，亦即键树的叶子总数；

Mk：所有长度为k的ACGT序列的组合数；

MN：在N个长为L的DNA序列中，出现的长度为k的子列的个数；

Node：键树所需的节点个数。

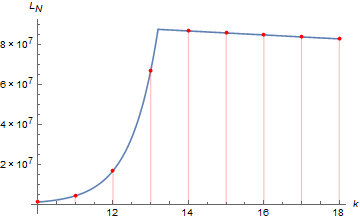
键树叶子个数分析：

最好条件下，叶子总数显然为1。

在最坏条件下，键树叶子总数Ln受到Mk和MN两个条件约束，在数值上取两者的最小值，亦即：

C:\Users\Administrator\Desktop\Eqn2.gif

取L和N为100与1,000,000，在Mathematica上绘制图像如下：



可见，当k=14时，Ln取最大值87,000,000。显然，随着N的减小，Ln将更早达到最大值，且最值将更小。

键树节点个数分析：

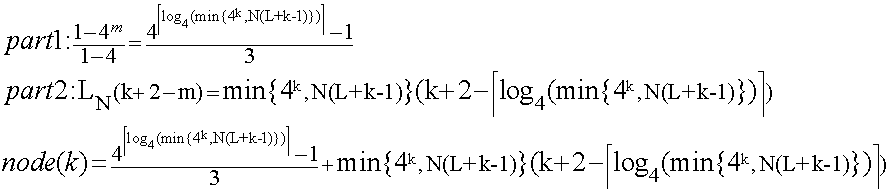
键树由一层根节点、k层中间节点、一层叶子组成，共k+2层。

在最优的情况下，键树只有一个叶子，故每层只有一个节点，节点个数Node=k+2。

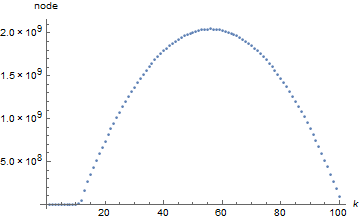
现在主要分析最坏的情况。最坏的情况下，键树的前半部分(part1)是一个完全四叉树，后半部分(part2)不再有分支，每层节点个数均为叶子的个数。记part1的层数为m，第m层(最后一层)节点数为t，则part2的层数为k+2-m。一方面，由完全四叉树可知，t=4m-1，另一方面，又有约束t<Ln。我们取Ln为最坏条件下的结果，则可以解得：



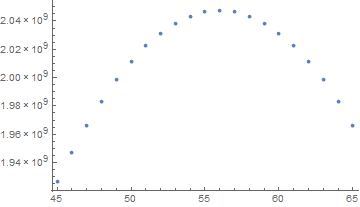
于是，part1和part2的节点个数与总节点个数分别为：



取L=100，N=1,000,000，在Mathematica上可制图如下：



顶点处放大如下：



在k=56时，node达到最大值2047369621。

叶子个数期望分析：

对于下面的分析是对k>=14而言的，对k<14的情况，不一定非要使用键树，所以在此不进行分析。

k>=14时，kMer的总数为1000000(101-k)条，要叶子个数为L，实际上相当于将这些kMer划分成L个集合。而将n个元素划分为m个集合的方法数，为S(n,m)，其中S代表第二类斯特灵数(Stirling Number of the Second Kind)，此处引文为：Abramowitz, M. and Stegun, I. A. (Eds.). "Stirling Numbers of the Second Kind." §24.1.4 in Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables, 9th printing. New York: Dover, pp. 824-825, 1972。而对n个元素总的划分方法数为贝尔数Bn(Bell Number)，显然有。

要求L的期望，实际上就是求S(n,m)对m的概率分布。而又有关系(里面的sigma应该替换成S的表示法)：

C:\Users\Administrator\AppData\Roaming\Tencent\Users\1020563055\QQ\WinTemp\RichOle\[JX$VLR][7AMD@)A3F5KF@Q.png

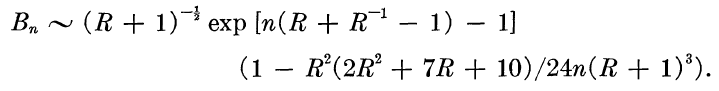
其中

C:\Users\Administrator\AppData\Roaming\Tencent\Users\1020563055\QQ\WinTemp\RichOle\VE35K66(~D$L(XBWD[V2A(5.png

(引文：L. H. Harper. “Stirling Behavior is Asymptotically Normal” Ann. Math. Statist. Volume 38, Number 2 (1967), 410-414)

令x=0，显然就得到了S(n,m)对m的期望值：Bn+1/Bn-1。

然而，对于较大的n，计算贝尔数的运算量极大，我们采用一个渐进表达式：



其中，R是方程R exp(R)=n的解，亦即W(n)，其中W为朗伯W函数(Lambert W-Function)。此处引文：C:\Users\Administrator\AppData\Roaming\Tencent\Users\1020563055\QQ\WinTemp\RichOle\7%[)0LUO6T)4$}5}X8X120P.png

这是一个非常好的近似估计，在n很小的时候就能达到较高的精度。比如对B5，实际值为52，通过该式的估计值为。

对14~100的k值，算得数值如下:

=======================列表======================

{14, 5.59913\*10^6},

{15, 5.53864\*10^6},

{16, 5.47812\*10^6},

{17, 5.41755\*10^6},

{18, 5.35694\*10^6},

{19, 5.29629\*10^6},

{20, 5.2356\*10^6},

{21, 5.17487\*10^6},

{22, 5.11409\*10^6},

{23, 5.05327\*10^6},

{24, 4.9924\*10^6},

{25, 4.93149\*10^6},

{26, 4.87054\*10^6},

{27, 4.80953\*10^6},

{28, 4.74848\*10^6},

{29, 4.68738\*10^6},

{30, 4.62623\*10^6},

{31, 4.56504\*10^6},

{32, 4.50379\*10^6},

{33, 4.44249\*10^6},

{34, 4.38114\*10^6},

{35, 4.31974\*10^6},

{36, 4.25828\*10^6},

{37, 4.19677\*10^6},

{38, 4.1352\*10^6},

{39, 4.07358\*10^6},

{40, 4.0119\*10^6},

{41, 3.95016\*10^6},

{42, 3.88836\*10^6},

{43, 3.8265\*10^6},

{44, 3.76458\*10^6},

{45, 3.70259\*10^6},

{46, 3.64054\*10^6},

{47, 3.57842\*10^6},

{48, 3.51624\*10^6},

{49, 3.45399\*10^6},

{50, 3.39167\*10^6},

{51, 3.32928\*10^6},

{52, 3.26681\*10^6},

{53, 3.20427\*10^6},

{54, 3.14166\*10^6},

{55, 3.07896\*10^6},

{56, 3.01619\*10^6},

{57, 2.95333\*10^6},

{58, 2.89039\*10^6},

{59, 2.82736\*10^6},

{60, 2.76424\*10^6},

{61, 2.70103\*10^6},

{62, 2.63773\*10^6},

{63, 2.57433\*10^6},

{64, 2.51084\*10^6},

{65, 2.44724\*10^6},

{66, 2.38353\*10^6},

{67, 2.31972\*10^6},

{68, 2.25579\*10^6},

{69, 2.19175\*10^6},

{70, 2.12759\*10^6},

{71, 2.0633\*10^6},

{72, 1.99888\*10^6},

{73, 1.93433\*10^6},

{74, 1.86964\*10^6},

{75, 1.80481\*10^6},

{76, 1.73982\*10^6},

{77, 1.67467\*10^6},

{78, 1.60936\*10^6},

{79, 1.54388\*10^6},

{80, 1.47821\*10^6},

{81, 1.41235\*10^6},

{82, 1.34629\*10^6},

{83, 1.28001\*10^6},

{84, 1.2135\*10^6},

{85, 1.14675\*10^6},

{86, 1.07974\*10^6},

{87, 1.01245\*10^6},

{88, 944850.},

{89, 876924.},

{90, 808638.},

{91, 739954.},

{92, 670826.},

{93, 601201.},

{94, 531005.},

{95, 460146.},

{96, 388498.},

{97, 315877.},

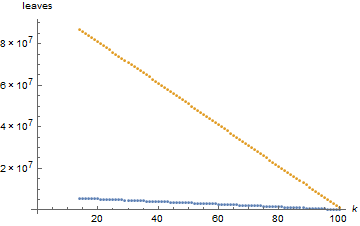
{98, 241999.},

{99, 166362.},

{100, 87846.6}

==========列表结束==================================

绘制图形如下：



可见，期望值远远小于估计的最大值，在最大值的6.4%到8.7%之间。而且k与期望叶子数具有良好的线性关系，用最小二乘法拟合得到，，可用于下一步的复杂度分析。

BUF\_LEN分析：

在第二步的过程中，即使在最极端情况，1000000\*100\*2\*sizeof(int)=0.8G<<8G，内存也满足要求，故而若BUF\_LEN取最小值1一定不会超出内存限制。

显然，在允许范围内，BUF\_LEN越大，文件写入的次数就越少，所以可以选取，其中leaves取4k(模型一)或叶子期望数(模型二)即可。对于1至100的k，算得列表：

{218750000, 54687500, 13671875, 3417968, 854492, 213623, 53405, 13351, 3337, 834, 208, 52, 13, 154, 156, 158, 160, 162, 164, 166, 168, 170, 172, 174, 176, 178, 181, 183, 186, 188, 191, 194, 196, 199, 202, 205, 208, 211, 215, 218, 222, 225, 229, 233, 237, 241, 246, 250, 255, 259, 264, 270, 275, 281, 287, 293, 299, 306, 313, 320, 327, 335, 344, 353, 362, 372, 382, 393, 405, 417, 430, 444, 459, 474, 491, 510, 529, 550, 573, 598, 625, 655, 688, 724, 764, 809, 860, 917, 982, 1058, 1146, 1250, 1374, 1526, 1717, 1961, 2287, 2741, 3422, 4553}