RECHERCHE DE MOTIFS RÉPÉTÉS DANS UN GÉNOME

Auteurs :

- Leo PERARD
- Salla DIAGNE

Listing des fichiers et répertoires du projet

- *bin/* : contient les fichiers sources compilés (.class)
- *donnees/* : contient des fichiers de tests au format FASTA
- *src/* : contient les fichiers JAVA du projet
- *test/* : contient les tests unitaires concernant le projet
- *strand_searching.jar* : jar contenant le programme principal

Fonctionnement du programme

```
DESCRIPTION : recherche de motifs repetes dans un genome
USAGE : java -jar strand_searching.jar filename [strand|N] [-comp|-rev|-revComp]*
--USING [-bf|-so|-kr|-kmp|-bm]*
    filename : le nom du fichier fasta ou se trouve le genome a etudier
    [strand|N] : permet de rechercher soit :
        * strand : une sequence dont les occurences seront recherchees dans le
genome
        * N : rechercher les occurences des mots de taille N
    [-comp|-rev|-revComp] : permettent de rechercher egalement pour le mot entre ou
les occurences des mots de taille N :
        * comp : le complementaire
        * rev : le reverse
        * revComp : le reverse-complementaire
        * dotplot : pour generer un dotplot comparant le genome a lui-meme
    [-bf|-so|-kr|-kmp|-bm] : permet de spécifier le ou les algos a rechercher parmi
:
        * bf : Brute-force
        * so : Shift-Or
        * kr : Karp-Rabin
        * kmp : Knutt-Morris-Pratt
        * bm : Boyer-Moore
Si aucun algorithme n'est specifie, l'algorithme de Boyer-Moore sera utilise.
EXEMPLE : java -jar strand_searching.jar donnees/simple.fasta TATA --WITH -revComp
-comp -rev --USING -kr -bf -so -bm -kmp
    Cet exemple affichera sur la sortie standard les occurences du mot TATA, de son
reverse, de son complementaire et de son
    reverse-complementaire dans le genome du fichier donnees/simple.fasta, en
utilisant les algorithme Karp-Rabin, Brute-Force,
    ShiftOr, Boyer-Moore et Knuth-Morris-Pratt
```

Exemples de résultats du programme

Recherche d'un motif et ses associés en particulier

```
$ java -jar strand_searching.jar donnees/exemple3.fasta GATA --WITH -comp -rev
-revComp --USING -bf -so -kr -kmp -bm

taille du genome : 1550
taille des motifs : 4

Algorithme naif (BruteForce)
GATA : [143, 173, 710, 796, 1021]
ATAG : [1022]
CTAT : []
TATC : [557, 1518]
1547 comparaisons pour chaque mot.
Temps d'execution : 6609009 nanosecondes.

Algorithme ShiftOr
GATA : [143, 173, 710, 796, 1021]
ATAG : [1022]
CTAT : []
TATC : [557, 1518]
1549 comparaisons pour chaque mot.
Temps d'execution : 17261901 nanosecondes.

Algorithme de Karp-Rabin
GATA : [143, 173, 710, 796, 1021]
ATAG : [1022]
CTAT : []
TATC : [557, 1518]
1547 comparaisons pour chaque mot.
Temps d'execution : 21542546 nanosecondes.

Algorithme de Knuth-Morris-Pratt
GATA : [143, 173, 710, 796, 1021]
ATAG : [1022]
CTAT : []
TATC : [557, 1518]
410 comparaisons pour chaque mot.
Temps d'execution : 3440243 nanosecondes.

Algorithme de Boyer-Moore
GATA : [143, 173, 710, 796, 1021]
ATAG : [1022]
CTAT : []
TATC : [557, 1518]
204 comparaisons pour chaque mot.
Temps d'execution : 2415388 nanosecondes.

$
```

Recherche de motifs d'une taille donnée avec génération de dotplot

```
$ java -jar strand_searching.jar donnees/exemple3.fasta 4 --WITH -comp -rev
-revComp -dotplot --USING -bm

Algorithme de Boyer-Moore
TTTT : [1456]
TTTG : [8, 181, 1425, 1457]
TTTA : [855, 961]
TTTC : [423, 594, 741, 846, 1546]
TTGT : [443, 1065, 1312, 1395, 1426]
TTGG : [0]
TTGA : [9, 182, 202, 467, 484, 657, 926, 997, 1121]
TTGC : [79, 1131]
TTAT : [85, 226, 556, 566, 1216]
TTAG : [93, 248, 446, 793, 838, 856, 1127, 1140]
TTAA : []
TTAC : [987]
TTCT : [595, 742, 1029, 1299]
TTCG : [195, 214, 424, 966, 1035, 1306]
TTCA : []
TTCC : []
TGTT : []
TGTG : [608, 687, 890, 956]
TGTA : [689, 750, 1313, 1396, 1427]
TGTC : [1066, 1079]
TGGT : [192, 257, 591, 745, 803, 958, 1055, 1062]
TGGG : [119, 140, 321, 570, 649, 777, 1092, 1145, 1188, 1224, 1445, 1478, 1491]
TGGA : [1, 67, 229, 233, 379, 631, 642, 677, 714, 892, 949, 1245, 1277, 1341, 1459,
1534]
TGGC : [16, 33, 477, 547, 728]
TGAT : [10, 308, 413, 605, 1122, 1487]
TGAG : [108, 260, 300, 327, 409, 658, 762, 825, 1084]
TGAA : []
TGAC : [391, 485, 756, 927, 998, 1008, 1048, 1160, 1170, 1194, 1219]
TGCT : [80, 461, 829, 860, 1237]
TGCG : [287, 702, 1528]
TGCA : [52, 662, 1058, 1262, 1320]
TGCC : []
TATT : []
TATG : [86, 227, 712, 1217]
TATA : []
TATC : [557, 1518]
TAGT : [253, 464, 806, 857, 1128, 1141, 1350]
TAGG : []
TAGA : [706, 794, 839, 1017]
TAGC : [94, 249, 291, 691, 1512]
TAAT : [45, 165, 536, 963, 1283, 1353]
TAAG : [72, 134, 472, 599, 832, 877, 1097, 1155, 1473]
TAAA : [433, 576, 818, 869, 1291]
TAAC : [112, 123, 145, 175, 265, 495, 513, 752, 1429, 1450, 1504]
TACT : []
```

```
TACG : [347, 517, 539, 897, 1381]
TACA : [48, 1229, 1240, 1252, 1331, 1398]
TACC : []
TCTT : []
TCTG : [389, 439, 603, 743, 748, 1006]
TCTA : [1153]
TCTC : [1030, 1286, 1300]
TCGT : [219, 430, 1068, 1076, 1081, 1501]
TCGG : []
TCGA : [58, 196, 967]
TCGC : [243, 1326, 1346, 1356]
TCAT : [638, 1137, 1204, 1207, 1288]
TCAG : [20, 1071, 1302, 1363]
TCAA : []
TCAC : [222, 272, 1411, 1538]
TCCT : [13, 344, 1003, 1032, 1543]
TCCG : [149, 372, 559, 848, 884]
TCCA : [683, 809]
TCCC : [1101, 1386]
GTTT : [7, 593, 845, 960, 1424]
GTTG : [255, 442, 466, 1064, 1090, 1130, 1143]
GTTA : [92, 445, 555, 837, 867, 1095]
GTTC : [194, 1298, 1305, 1384]
GTGT : [688, 835, 889, 1078]
GTGG : [1054, 1187, 1477]
GTGA : [107, 259, 307, 408, 412, 609, 696, 1047, 1083, 1159, 1169, 1376, 1495]
GTGC : [40, 460, 520, 661, 828, 859, 1057, 1236, 1527]
GTAT : [1517]
GTAG : [290, 361, 542, 690, 705, 805, 1314, 1511]
GTAA : []
GTAC : [490, 896, 1251, 1397]
GTCT : [388, 602, 747, 994]
GTCG : [57, 218, 242, 429, 1067, 1080, 1440, 1500]
GTCA : [221, 637, 1070, 1198, 1410]
GTCC : []
GGTT : []
GGTG : [106, 258, 306, 545, 589, 695, 947, 1053, 1056, 1158, 1168, 1186, 1275, 1375,
1443, 1476, 1494, 1526]
GGTA : [121, 263, 489, 534, 804, 1250, 1448, 1510]
GGTC : [217, 428, 636, 746, 993]
GGGT : [105, 120, 305, 427, 635, 843, 1093, 1447, 1493]
GGGG : [160, 932]
GGGA : [141, 152, 322, 350, 364, 449, 650, 779, 1038, 1190, 1479]
GGGC : [161, 571, 933, 1025, 1146, 1225, 1390]
GGAT : [23, 68, 142, 172, 230, 790, 1191, 1309, 1360, 1535]
GGAG : [2, 351, 395, 632, 651, 715, 780, 893, 950, 1039, 1179, 1278, 1460]
GGAA : [922]
GGAC : [101, 234, 323, 380, 1246, 1480]
GGCT : [17, 162, 211, 269, 420, 510, 621, 908, 1026, 1226, 1317, 1531]
GGCG : [34, 37, 98, 208, 572, 735, 765]
GGCA : [279, 354, 478, 548]
GGCC : []
GATT : []
GATG : [24, 231, 285, 414, 606, 823, 1087, 1192, 1485]
GATA : [143, 173, 710, 796, 1021]
GATC : [11, 309, 1123, 1361, 1536]
GAGT : [5, 109, 410, 659, 826, 894, 1045, 1422]
```

```
GAGG : [352, 1180]
GAGA : [3, 301, 328, 652, 708, 1019, 1085, 1420]
GAGC : [60, 75, 396, 781, 951, 1040, 1110, 1258, 1279, 1461]
GAAT : [64, 366, 469, 563, 679, 923, 1343, 1378]
GAAG : [89, 417, 451, 666, 864, 969, 979, 1183, 1336, 1437, 1497, 1523]
GAAA : [154, 198, 204, 384, 501, 611, 644, 672, 698, 770, 912]
GAAC : [27, 184, 454, 580, 719, 982]
GACT : [137, 324, 341, 654, 738, 1161]
GACG : [102, 381, 392, 486, 928, 1195, 1247]
GACA : [330, 757, 999, 1009, 1049, 1171, 1481]
GACC : [296]
GCTT : []
GCTG : [31, 475, 861, 909, 1260, 1318, 1339, 1532]
GCTA : [163, 251, 462, 511, 830, 1227, 1238, 1281, 1348]
GCTC : [18, 81, 270, 622, 1074]
GCGT : [38, 240, 288, 406, 553, 573, 703, 774, 887]
GCGG : [209]
GCGA : [62, 736, 768, 977, 1271]
GCGC : [584, 766, 1112]
GCAT : [189, 245, 873, 953, 1059, 1366]
GCAG : [355, 358, 526, 586, 663, 1042]
GCAA : [972]
GCAC : [479, 880, 939]
GCCT : [42, 1328, 1392, 1470]
GCCG : [293, 403, 529, 814, 901, 1165, 1370, 1467, 1514]
GCCA : [314, 505, 522, 1133, 1148, 1463]
GCCC : [129, 336, 615, 730, 851, 935, 1211, 1406]
ATTT : [179]
ATTG : [201, 568, 640, 925, 1311, 1489]
ATTA : [70, 247, 565, 792, 875, 1139]
ATTC : [681, 965]
ATGT : [607, 686, 955, 1088]
ATGG : []
ATGA : [25, 87, 415, 824, 1193, 1218, 1334, 1486]
ATGC : [51, 286, 701, 1209, 1368]
ATAT : [711]
ATAG : [1022]
ATAA : []
ATAC : [47, 167, 538, 797, 1380]
ATCT : [368, 1124, 1285]
ATCG : [310, 1345, 1355, 1519]
ATCA : [1136, 1203, 1206, 1362, 1537]
ATCC : []
AGTT : [1142]
AGTG : [411, 459, 660, 675, 726, 827, 834, 858, 1046]
AGTA : [110, 360, 895, 1351]
AGTC : [56, 387, 601, 807, 1099, 1439, 1499]
AGGT : [262, 544, 588, 992, 1052, 1157, 1185, 1274, 1475, 1509, 1525]
AGGG : [304, 363, 448, 634, 842, 1024]
AGGA : [22, 670, 717, 789, 921, 1181]
AGGC : [207, 278, 353, 419, 734, 764, 907, 1316]
AGAT : [709, 795, 1020, 1086, 1484]
AGAG : [4, 74, 302, 668, 707, 840, 1018, 1044, 1257, 1421]
AGAA : [579]
AGAC : [136, 329, 340, 653, 1266]
AGCT : [76, 250, 474, 1073, 1259, 1280, 1338]
AGCG : [61, 95, 552, 692, 773, 944, 1111]
```

```
AGCA : [357, 397, 525, 782, 872, 879, 952, 971, 1041, 1365]
AGCC : []
AATT : [964]
AATG : [65, 377, 700, 1243]
AATA : [46, 166, 470, 537, 1379]
AATC : [367, 1202, 1284, 1344, 1354]
AAGT : [55, 90, 386, 458, 600, 674, 833, 865, 1098, 1438, 1498]
AAGG : []
AAGA : [73, 135, 452, 578, 667, 980, 1256, 1265]
AAGC : [473, 503, 551, 613, 772, 871, 878, 943, 970, 1337]
AAAT : [199, 699, 1201]
AAAG : [205, 385, 502, 577, 612, 673, 771, 870, 919, 1255]
AAAA : []
AAAC : [155, 435, 645, 785, 819, 913, 1174, 1293]
AACT : [146, 436, 514, 646, 753, 914, 1323]
AACG : [28, 266, 282, 400, 581, 820, 974, 1107]
AACA : [113, 176, 455, 720, 786, 1430, 1505]
AACC : []
ACTT : []
ACTG : [138, 319, 325, 647, 754, 760, 1162]
ACTA : [515]
ACTC : [147, 342, 437, 882, 915, 1151, 1324]
ACGT : [116, 518, 540, 1196, 1234, 1382]
ACGG : []
ACGA : [283, 382, 821, 1108, 1418]
ACGC : [29, 401, 582, 812, 975]
ACAT : [49, 177, 1000, 1332]
ACAG : [787, 1050, 1482]
ACAA : [456, 1506]
ACAC : [114, 317, 331, 721, 758, 1230, 1232, 1399, 1401, 1413, 1431]
ACCT : [125, 297, 481, 492, 984, 1221, 1452, 1540]
ACCG : [157, 169, 186, 627, 1176, 1268, 1295, 1403]
ACCA : [274, 497, 723, 989, 1415]
ACCC : [236, 799, 1013, 1117, 1433]
CTTT : [422, 740, 854, 1454, 1545]
CTTG : [78, 483, 656, 996, 1120, 1394]
CTTA : [84, 225, 597, 986, 1126, 1215]
CTTC : [213, 370, 1028, 1034]
CTGT : [440, 749]
CTGG : []
CTGA : [299, 326, 390, 604, 755, 761, 862, 910, 1007]
CTGC : [127, 1163, 1261, 1319]
CTAT : []
CTAG : [252, 463, 1016, 1349]
CTAA : []
CTAC : [346, 516, 1228, 1239, 1330]
CTCT : [82, 438, 1005, 1152]
CTCG : [1075, 1325]
CTCA : [19, 271, 623, 916, 1287, 1301]
CTCC : []
CGTT : []
CGTG : [39, 117, 407, 519, 629, 775, 888, 1077, 1082, 1235]
CGTA : [289, 431, 541, 574, 704, 816, 1502, 1516]
CGTC : [220, 241, 1069, 1197, 1409]
CGGT : [216, 488, 533, 694, 946, 1167, 1249, 1374, 1442]
CGGG : [104, 151, 159, 349, 426, 930, 1037, 1389]
CGGA : [100, 171, 394, 561, 1178, 1308, 1359, 1521]
```

```
CGGC : [36, 97, 210, 268, 312, 334, 509, 620, 899, 1530]
CGAT : [284, 822]
CGAG : [59, 1109, 1272, 1419]
CGAA : []
CGAC : [295, 737]
CGCT : [30, 1347]
CGCG : [239, 405, 531, 583, 767, 886, 976, 1270, 1357, 1372]
CGCA : [188, 244, 374, 585, 903, 938, 1104, 1113]
CGCC : []
CATT : []
CATG : [50, 190, 685, 954, 1060, 1208, 1333, 1367]
CATA : [132, 1289]
CATC : [1001, 1135, 1205]
CAGT : [359, 725, 1303]
CAGG : []
CAGA : [339, 499, 664, 1043, 1483]
CAGC : [356, 524, 527, 1072, 1364, 1465]
CAAT : [376, 1242]
CAAG : [54, 276, 457, 550, 905, 942, 1264, 1507]
CAAA : [784, 918, 1173, 1200, 1254]
CAAC : [281, 399, 625, 973, 1011, 1106, 1115, 1322]
CACT : [223, 318, 759, 881, 1150]
CACG : [115, 332, 507, 618, 811, 1233, 1417]
CACA : [316, 940, 1231, 1400, 1412]
CACC : []
CCTT : []
CCTG : [14, 126, 298, 801, 1222]
CCTA : [43, 345, 493, 1015, 1329, 1471]
CCTC : [1004, 1541]
CCGT : [628, 815, 1296, 1408, 1515]
CCGG : []
CCGA : [294, 1435]
CCGC : [187, 238, 373, 404, 530, 849, 885, 902, 937, 1103, 1269, 1371, 1404, 1468]
CCAT : [131, 684, 1134]
CCAG : [338, 498, 523, 724, 732, 990, 1464]
CCAA : []
CCAC : [315, 506, 617, 810, 1149, 1416]
CCCT : [800, 852, 1014, 1118, 1213]
CCCG : [237, 936, 1102, 1387, 1407, 1434]
CCCA : [130, 337, 616, 731]
CCCC : [1212]
160 comparaisons pour chaque mot.
Temps d'execution : 33558598 nanosecondes.

Generation du dotplot...
Dotplot genere avec succes (fichier dotplot.jpg)

$
```