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)
- *conf/* : contient les fichiers de configuration de l'alphabet et des appariemments
- *donnees/* : contient des fichiers de tests au format FASTA
- *lib/* : contient les librairies dont a besoin le projet (typiquement des librairies de tests unitaires)
- *src/* : contient les fichiers JAVA du projet
- *test/* : contient les tests unitaires concernant le projet
- *strand_searching.jar* : jar contenant le programme principal

Algorithmes de recherche implémentés

- Brute-Force
- Shift-Or
- Karp-Rabin
- Knuth-Morris-Pratt
- Boyer-Moore

Fonctionnement du programme

```
DESCRIPTION : recherche de motifs repetes dans un genome
USAGE : java -jar strand_searching.jar confFilename genomeFilename [strand|N]
--WITH [-comp|-rev|-revComp]* --USING [-bf|-so|-kr|-kmp|-bm]* [--DOTPLOT]
    confFilename : le nom du fichier de configuration des paires (conf/acgt.conf ou
conf/a)
    genomeFilename : 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 Knuth-Morris-Pratt 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 conf/acgt.conf strand_searching.jar donnees/exemple3.fasta GATA --WITH
-comp -rev -revComp --USING -bf -kr -so -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]
8 occurences trouvees au total.
1547 comparaisons pour chaque mot.
Temps d'execution : 6371268 nanosecondes.

Algorithme de Karp-Rabin
GATA : [143, 173, 710, 796, 1021]
ATAG : [1022]
CTAT : []
TATC : [557, 1518]
8 occurences trouvees au total.
1547 comparaisons pour chaque mot.
Temps d'execution : 25803073 nanosecondes.

Algorithme ShiftOr
GATA : [143, 173, 710, 796, 1021]
ATAG : [1022]
CTAT : []
TATC : [557, 1518]
8 occurences trouvees au total.
1549 comparaisons pour chaque mot.
Temps d'execution : 18965818 nanosecondes.
```

```
Algorithme de Knuth-Morris-Pratt
GATA : [143, 173, 710, 796, 1021]
ATAG : [1022]
CTAT : []
TATC : [557, 1518]
8 occurences trouvees au total.
410 comparaisons pour chaque mot.
Temps d'execution : 1285429 nanosecondes.

Algorithme de Boyer-Moore
GATA : [143, 173, 710, 796, 1021]
ATAG : [1022]
CTAT : []
TATC : [557, 1518]
8 occurences trouvees au total.
204 comparaisons pour chaque mot.
Temps d'execution : 3946804 nanosecondes.

$
```

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

```
$ java -jar strand_searching.jar conf/acgu.conf donnees/exemple3.fasta 4 --USING
-kmp --DOTPLOT

taille du genome : 1550
taille des motifs : 2

Algorithme de Knuth-Morris-Pratt
UU : [0, 8, 9, 71, 79, 85, 93, 180, 181, 182, 195, 202, 214, 226, 248, 256, 371,
423, 424, 443, 446, 467, 484, 556, 566, 569, 594, 595, 598, 641, 657, 682, 741, 742,
793, 838, 846, 847, 855, 856, 868, 876, 926, 961, 962, 966, 987, 997, 1029, 1035,
1065, 1091, 1096, 1121, 1127, 1131, 1140, 1144, 1216, 1299, 1306, 1312, 1385, 1395,
1425, 1426, 1455, 1456, 1457, 1458, 1490, 1546, 1547]
UG : [1, 10, 16, 26, 33, 41, 52, 67, 80, 88, 108, 119, 128, 140, 183, 192, 203, 229,
233, 257, 260, 287, 300, 308, 321, 327, 379, 391, 409, 413, 416, 441, 444, 461, 468,
477, 485, 521, 547, 570, 591, 605, 608, 610, 631, 642, 649, 658, 662, 677, 687, 689,
697, 702, 714, 728, 745, 750, 756, 762, 777, 803, 825, 829, 836, 860, 863, 890, 892,
911, 927, 949, 956, 958, 998, 1008, 1048, 1055, 1058, 1062, 1066, 1079, 1084, 1089,
1092, 1122, 1132, 1145, 1160, 1164, 1170, 1188, 1194, 1210, 1219, 1224, 1237, 1245,
1262, 1277, 1313, 1320, 1335, 1341, 1369, 1377, 1396, 1427, 1445, 1459, 1478, 1487,
1491, 1496, 1528, 1534]
UA : [45, 48, 72, 86, 94, 112, 123, 134, 145, 165, 168, 175, 227, 249, 253, 265,
291, 347, 362, 433, 447, 464, 472, 491, 495, 513, 517, 536, 539, 543, 557, 567, 576,
599, 691, 706, 712, 752, 794, 798, 806, 818, 832, 839, 857, 869, 877, 897, 963, 988,
1017, 1023, 1097, 1128, 1141, 1155, 1217, 1229, 1240, 1252, 1283, 1291, 1315, 1331,
1350, 1353, 1381, 1398, 1429, 1450, 1473, 1504, 1512, 1518]
UC : [13, 20, 58, 83, 149, 196, 215, 219, 222, 243, 272, 311, 344, 369, 372, 389,
425, 430, 439, 559, 596, 603, 624, 638, 683, 743, 748, 809, 848, 884, 917, 967, 995,
1003, 1006, 1030, 1032, 1036, 1068, 1071, 1076, 1081, 1101, 1125, 1137, 1153, 1199,
1204, 1207, 1286, 1288, 1300, 1302, 1307, 1326, 1346, 1356, 1363, 1386, 1411, 1441,
1501, 1520, 1538, 1543, 1548]
GU : [7, 40, 57, 92, 107, 111, 118, 122, 194, 218, 221, 242, 255, 259, 264, 290,
```

307, 361, 388, 408, 412, 429, 432, 442, 445, 460, 466, 490, 520, 535, 542, 546, 555,
575, 590, 593, 602, 609, 630, 637, 661, 676, 688, 690, 696, 705, 727, 747, 751, 776,
805, 808, 817, 828, 835, 837, 845, 859, 867, 889, 891, 896, 948, 957, 960, 994,
1047, 1054, 1057, 1064, 1067, 1070, 1078, 1080, 1083, 1090, 1095, 1100, 1130, 1143,
1159, 1169, 1187, 1198, 1236, 1251, 1276, 1298, 1305, 1314, 1352, 1376, 1384, 1397,
1410, 1424, 1428, 1440, 1444, 1449, 1477, 1495, 1500, 1503, 1511, 1517, 1527]
GG : [2, 17, 23, 34, 37, 68, 98, 101, 105, 106, 120, 121, 141, 142, 152, 153, 160,
161, 162, 172, 193, 208, 211, 217, 230, 234, 258, 263, 269, 279, 305, 306, 313, 322,
323, 335, 350, 351, 354, 364, 365, 380, 395, 420, 427, 428, 449, 450, 478, 489, 510,
534, 545, 548, 562, 571, 572, 589, 592, 621, 632, 635, 636, 643, 650, 651, 671, 678,
695, 715, 718, 729, 735, 746, 765, 778, 779, 780, 790, 804, 843, 844, 893, 900, 908,
922, 931, 932, 933, 934, 947, 950, 959, 993, 1025, 1026, 1038, 1039, 1053, 1056,
1063, 1093, 1094, 1146, 1147, 1158, 1168, 1179, 1182, 1186, 1189, 1190, 1191, 1225,
1226, 1246, 1250, 1275, 1278, 1309, 1317, 1342, 1360, 1375, 1390, 1391, 1443, 1446,
1447, 1448, 1460, 1476, 1479, 1480, 1492, 1493, 1494, 1510, 1522, 1526, 1531, 1535]
GA : [3, 5, 11, 24, 27, 60, 64, 69, 75, 89, 102, 109, 137, 143, 154, 173, 184, 198,
204, 231, 235, 261, 285, 296, 301, 303, 309, 324, 328, 330, 341, 352, 366, 381, 384,
392, 396, 410, 414, 417, 451, 454, 469, 486, 501, 563, 580, 606, 611, 633, 644, 652,
654, 659, 666, 669, 672, 679, 698, 708, 710, 716, 719, 738, 757, 763, 770, 781, 791,
796, 823, 826, 841, 864, 894, 912, 923, 928, 951, 969, 979, 982, 999, 1009, 1019,
1021, 1040, 1045, 1049, 1085, 1087, 1110, 1123, 1161, 1171, 1180, 1183, 1192, 1195,
1220, 1247, 1258, 1267, 1273, 1279, 1310, 1336, 1343, 1361, 1378, 1420, 1422, 1437,
1461, 1481, 1485, 1488, 1497, 1523, 1536]
GC : [18, 31, 35, 38, 42, 53, 62, 77, 81, 96, 99, 129, 163, 189, 209, 212, 240, 245,
251, 270, 280, 288, 293, 314, 336, 355, 358, 375, 398, 403, 406, 421, 462, 475, 479,
505, 511, 522, 526, 529, 532, 549, 553, 573, 584, 586, 615, 622, 663, 693, 703, 730,
736, 766, 768, 774, 783, 814, 830, 851, 861, 873, 880, 887, 901, 904, 909, 935, 939,
945, 953, 972, 977, 1027, 1042, 1059, 1074, 1105, 1112, 1114, 1133, 1148, 1165,
1211, 1227, 1238, 1260, 1263, 1271, 1281, 1318, 1321, 1328, 1339, 1348, 1358, 1366,
1370, 1373, 1392, 1406, 1463, 1467, 1470, 1514, 1529, 1532]
AU : [12, 25, 47, 51, 66, 70, 87, 133, 144, 167, 174, 179, 191, 201, 228, 232, 247,
286, 310, 368, 378, 415, 471, 538, 558, 565, 568, 607, 640, 681, 686, 701, 711, 713,
792, 797, 824, 875, 925, 955, 965, 1002, 1022, 1061, 1088, 1124, 1136, 1139, 1193,
1203, 1206, 1209, 1218, 1244, 1285, 1290, 1311, 1334, 1345, 1355, 1362, 1368, 1380,
1486, 1489, 1519, 1537]
AG : [4, 6, 22, 56, 61, 74, 76, 91, 95, 110, 136, 207, 250, 254, 262, 278, 292, 302,
304, 329, 340, 353, 357, 360, 363, 387, 397, 411, 419, 448, 453, 459, 465, 474, 500,
504, 525, 528, 544, 552, 579, 588, 601, 614, 634, 653, 660, 665, 668, 670, 675, 692,
707, 709, 717, 726, 734, 764, 773, 782, 789, 795, 807, 827, 834, 840, 842, 858, 866,
872, 879, 895, 907, 921, 944, 952, 971, 981, 992, 1018, 1020, 1024, 1041, 1044,
1046, 1052, 1073, 1086, 1099, 1111, 1129, 1142, 1157, 1181, 1185, 1257, 1259, 1266,
1274, 1280, 1304, 1316, 1338, 1351, 1365, 1421, 1423, 1439, 1462, 1466, 1475, 1484,
1499, 1509, 1513, 1525]
AA : [28, 46, 55, 65, 73, 90, 113, 124, 135, 146, 155, 156, 166, 176, 185, 199, 200,
205, 206, 266, 277, 282, 367, 377, 385, 386, 400, 418, 434, 435, 436, 452, 455, 458,
470, 473, 496, 502, 503, 514, 537, 551, 564, 577, 578, 581, 600, 612, 613, 626, 645,
646, 667, 673, 674, 680, 699, 700, 720, 753, 771, 772, 785, 786, 819, 820, 833, 865,
870, 871, 878, 906, 913, 914, 919, 920, 924, 943, 964, 970, 974, 980, 983, 1012,
1098, 1107, 1116, 1156, 1174, 1175, 1184, 1201, 1202, 1243, 1255, 1256, 1265, 1284,
1292, 1293, 1294, 1323, 1337, 1344, 1354, 1379, 1430, 1438, 1451, 1474, 1498, 1505,
1508, 1524]
AC : [29, 49, 103, 114, 116, 125, 138, 147, 157, 169, 177, 186, 224, 236, 267, 274,
283, 297, 317, 319, 325, 331, 333, 342, 348, 382, 393, 401, 437, 456, 481, 487, 492,
497, 508, 515, 518, 540, 582, 619, 627, 647, 655, 721, 723, 739, 754, 758, 760, 787,
799, 812, 821, 882, 898, 915, 929, 941, 975, 984, 989, 1000, 1010, 1013, 1050, 1108,
1117, 1151, 1162, 1172, 1176, 1196, 1221, 1230, 1232, 1234, 1241, 1248, 1253, 1268,
1295, 1324, 1332, 1382, 1399, 1401, 1403, 1413, 1415, 1418, 1431, 1433, 1452, 1482,

```
1506, 1540]
CU : [15, 19, 32, 44, 78, 82, 84, 127, 139, 148, 164, 213, 225, 252, 271, 299, 320,
326, 343, 346, 370, 390, 422, 438, 440, 463, 476, 483, 494, 512, 516, 597, 604, 623,
648, 656, 740, 744, 749, 755, 761, 802, 831, 854, 862, 883, 910, 916, 986, 996,
1005, 1007, 1016, 1028, 1031, 1034, 1075, 1120, 1126, 1152, 1154, 1163, 1215, 1223,
1228, 1239, 1261, 1282, 1287, 1301, 1319, 1325, 1330, 1340, 1349, 1394, 1454, 1472,
1533, 1542, 1545]
CG : [30, 36, 39, 59, 63, 97, 100, 104, 117, 151, 159, 171, 188, 197, 210, 216, 220,
239, 241, 244, 268, 284, 289, 295, 312, 334, 349, 374, 383, 394, 402, 405, 407, 426,
431, 488, 509, 519, 531, 533, 541, 554, 561, 574, 583, 585, 620, 629, 694, 704, 737,
767, 769, 775, 813, 816, 822, 850, 886, 888, 899, 903, 930, 938, 946, 968, 976, 978,
1037, 1069, 1077, 1082, 1104, 1109, 1113, 1167, 1178, 1197, 1235, 1249, 1270, 1272,
1297, 1308, 1327, 1347, 1357, 1359, 1372, 1374, 1383, 1389, 1405, 1409, 1419, 1436,
1442, 1469, 1502, 1516, 1521, 1530]
CA : [21, 50, 54, 115, 132, 178, 190, 223, 246, 273, 276, 281, 316, 318, 332, 339,
356, 359, 376, 399, 457, 480, 499, 507, 524, 527, 550, 587, 618, 625, 639, 664, 685,
722, 725, 733, 759, 784, 788, 811, 874, 881, 905, 918, 940, 942, 954, 973, 991,
1001, 1011, 1043, 1051, 1060, 1072, 1106, 1115, 1135, 1138, 1150, 1173, 1200, 1205,
1208, 1231, 1233, 1242, 1254, 1264, 1289, 1303, 1322, 1333, 1364, 1367, 1400, 1402,
1412, 1414, 1417, 1432, 1465, 1483, 1507, 1539]
CC : [14, 43, 126, 130, 131, 150, 158, 170, 187, 237, 238, 275, 294, 298, 315, 337,
338, 345, 373, 404, 482, 493, 498, 506, 523, 530, 560, 616, 617, 628, 684, 724, 731,
732, 800, 801, 810, 815, 849, 852, 853, 885, 902, 936, 937, 985, 990, 1004, 1014,
1015, 1033, 1102, 1103, 1118, 1119, 1134, 1149, 1166, 1177, 1212, 1213, 1214, 1222,
1269, 1296, 1329, 1371, 1387, 1388, 1393, 1404, 1407, 1408, 1416, 1434, 1435, 1453,
1464, 1468, 1471, 1515, 1541, 1544]
1549 occurences trouvees au total.
434 comparaisons pour chaque mot.
Temps d'execution : 11098554 nanosecondes.

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

$
```