

## Introduction to Bioinformatics

Prof. Dr. Nizamettin AYDIN

### Local Multiple Sequence Alignment Sequence File Formats

[naydin@yildiz.edu.tr](mailto:naydin@yildiz.edu.tr)

<http://www3.yildiz.edu.tr/~naydin>

## Localized Alignments

- Just like with pairwise alignments, we may not be interested in the global alignment of multiple sequences, but rather only specific regions that are conserved.
- Local Alignment of **MSAs** are important:
  - Given regions of genomic DNA occurring upstream or before a certain gene, there might be sequences where transcription factors bind to the DNA so that the gene can be transcribed.
  - Thus, if we are interested in determining if there is any signal in the regions upstream of a certain family of genes across several different organisms, it would be important to only find the conserved region, and not try to align all of the genomic DNA.
  - Localized alignments of protein sequences can yield information about conserved domains found in otherwise unrelated proteins.

1

2

## Approaches to Local Alignment

- Profile Analysis
- Block Analysis
- Pattern-searching or statistical methods

## Profile Analysis...

- Profiles are found by first multiply aligning the sequences, determining which regions are the most highly conserved,
- and then creating a scoring matrix for the alignment of the highly conserved region.
- Profile is composed of:
  - **Columns:**
    - one for each residue;
    - columns for insertions and deletions as well
  - **Rows:**
    - one for each position in the conserved region or motif

3

4

## ...Profile Analysis

- Profiles describe a MSA by a scoring matrix:

```
      1 2 3 4
      5 6 7 8
      9 10 11 12
      13 14 15 16
      17 18 19 20
      21 22 23 24
      25 26 27 28
      29 30 31 32
      33 34 35 36
      37 38 39 40
      41 42 43 44
      45 46 47 48
      49 50 51 52
      53 54 55 56
      57 58 59 60
      61 62 63 64
      65 66 67 68
      69 70 71 72
      73 74 75 76
      77 78 79 80
      81 82 83 84
      85 86 87 88
      89 90 91 92
      93 94 95 96
      97 98 99 100
      101 102 103 104
      105 106 107 108
      109 110 111 112
      113 114 115 116
      117 118 119 120
      121 122 123 124
      125 126 127 128
      129 130 131 132
      133 134 135 136
      137 138 139 140
      141 142 143 144
      145 146 147 148
      149 150 151 152
      153 154 155 156
      157 158 159 160
      161 162 163 164
      165 166 167 168
      169 170 171 172
      173 174 175 176
      177 178 179 180
      181 182 183 184
      185 186 187 188
      189 190 191 192
      193 194 195 196
      197 198 199 200
      201 202 203 204
      205 206 207 208
      209 210 211 212
      213 214 215 216
      217 218 219 220
      221 222 223 224
      225 226 227 228
      229 230 231 232
      233 234 235 236
      237 238 239 240
      241 242 243 244
      245 246 247 248
      249 250 251 252
      253 254 255 256
      257 258 259 260
      261 262 263 264
      265 266 267 268
      269 270 271 272
      273 274 275 276
      277 278 279 280
      281 282 283 284
      285 286 287 288
      289 290 291 292
      293 294 295 296
      297 298 299 300
      301 302 303 304
      305 306 307 308
      309 310 311 312
      313 314 315 316
      317 318 319 320
      321 322 323 324
      325 326 327 328
      329 330 331 332
      333 334 335 336
      337 338 339 340
      341 342 343 344
      345 346 347 348
      349 350 351 352
      353 354 355 356
      357 358 359 360
      361 362 363 364
      365 366 367 368
      369 370 371 372
      373 374 375 376
      377 378 379 380
      381 382 383 384
      385 386 387 388
      389 390 391 392
      393 394 395 396
      397 398 399 400
      401 402 403 404
      405 406 407 408
      409 410 411 412
      413 414 415 416
      417 418 419 420
      421 422 423 424
      425 426 427 428
      429 430 431 432
      433 434 435 436
      437 438 439 440
      441 442 443 444
      445 446 447 448
      449 450 451 452
      453 454 455 456
      457 458 459 460
      461 462 463 464
      465 466 467 468
      469 470 471 472
      473 474 475 476
      477 478 479 480
      481 482 483 484
      485 486 487 488
      489 490 491 492
      493 494 495 496
      497 498 499 500
      501 502 503 504
      505 506 507 508
      509 510 511 512
      513 514 515 516
      517 518 519 520
      521 522 523 524
      525 526 527 528
      529 530 531 532
      533 534 535 536
      537 538 539 540
      541 542 543 544
      545 546 547 548
      549 550 551 552
      553 554 555 556
      557 558 559 560
      561 562 563 564
      565 566 567 568
      569 570 571 572
      573 574 575 576
      577 578 579 580
      581 582 583 584
      585 586 587 588
      589 590 591 592
      593 594 595 596
      597 598 599 600
      601 602 603 604
      605 606 607 608
      609 610 611 612
      613 614 615 616
      617 618 619 620
      621 622 623 624
      625 626 627 628
      629 630 631 632
      633 634 635 636
      637 638 639 640
      641 642 643 644
      645 646 647 648
      649 650 651 652
      653 654 655 656
      657 658 659 660
      661 662 663 664
      665 666 667 668
      669 670 671 672
      673 674 675 676
      677 678 679 680
      681 682 683 684
      685 686 687 688
      689 690 691 692
      693 694 695 696
      697 698 699 700
      701 702 703 704
      705 706 707 708
      709 710 711 712
      713 714 715 716
      717 718 719 720
      721 722 723 724
      725 726 727 728
      729 730 731 732
      733 734 735 736
      737 738 739 740
      741 742 743 744
      745 746 747 748
      749 750 751 752
      753 754 755 756
      757 758 759 760
      761 762 763 764
      765 766 767 768
      769 770 771 772
      773 774 775 776
      777 778 779 780
      781 782 783 784
      785 786 787 788
      789 790 791 792
      793 794 795 796
      797 798 799 800
      801 802 803 804
      805 806 807 808
      809 810 811 812
      813 814 815 816
      817 818 819 820
      821 822 823 824
      825 826 827 828
      829 830 831 832
      833 834 835 836
      837 838 839 840
      841 842 843 844
      845 846 847 848
      849 850 851 852
      853 854 855 856
      857 858 859 860
      861 862 863 864
      865 866 867 868
      869 870 871 872
      873 874 875 876
      877 878 879 880
      881 882 883 884
      885 886 887 888
      889 890 891 892
      893 894 895 896
      897 898 899 900
      901 902 903 904
      905 906 907 908
      909 910 911 912
      913 914 915 916
      917 918 919 920
      921 922 923 924
      925 926 927 928
      929 930 931 932
      933 934 935 936
      937 938 939 940
      941 942 943 944
      945 946 947 948
      949 950 951 952
      953 954 955 956
      957 958 959 960
      961 962 963 964
      965 966 967 968
      969 970 971 972
      973 974 975 976
      977 978 979 980
      981 982 983 984
      985 986 987 988
      989 990 991 992
      993 994 995 996
      997 998 999 1000
      1001 1002 1003 1004
      1005 1006 1007 1008
      1009 1010 1011 1012
      1013 1014 1015 1016
      1017 1018 1019 1020
      1021 1022 1023 1024
      1025 1026 1027 1028
      1029 1030 1031 1032
      1033 1034 1035 1036
      1037 1038 1039 1040
      1041 1042 1043 1044
      1045 1046 1047 1048
      1049 1050 1051 1052
      1053 1054 1055 1056
      1057 1058 1059 1060
      1061 1062 1063 1064
      1065 1066 1067 1068
      1069 1070 1071 1072
      1073 1074 1075 1076
      1077 1078 1079 1080
      1081 1082 1083 1084
      1085 1086 1087 1088
      1089 1090 1091 1092
      1093 1094 1095 1096
      1097 1098 1099 1100
      1101 1102 1103 1104
      1105 1106 1107 1108
      1109 1110 1111 1112
      1113 1114 1115 1116
      1117 1118 1119 1120
      1121 1122 1123 1124
      1125 1126 1127 1128
      1129 1130 1131 1132
      1133 1134 1135 1136
      1137 1138 1139 1140
      1141 1142 1143 1144
      1145 1146 1147 1148
      1149 1150 1151 1152
      1153 1154 1155 1156
      1157 1158 1159 1160
      1161 1162 1163 1164
      1165 1166 1167 1168
      1169 1170 1171 1172
      1173 1174 1175 1176
      1177 1178 1179 1180
      1181 1182 1183 1184
      1185 1186 1187 1188
      1189 1190 1191 1192
      1193 1194 1195 1196
      1197 1198 1199 1200
      1201 1202 1203 1204
      1205 1206 1207 1208
      1209 1210 1211 1212
      1213 1214 1215 1216
      1217 1218 1219 1220
      1221 1222 1223 1224
      1225 1226 1227 1228
      1229 1230 1231 1232
      1233 1234 1235 1236
      1237 1238 1239 1240
      1241 1242 1243 1244
      1245 1246 1247 1248
      1249 1250 1251 1252
      1253 1254 1255 1256
      1257 1258 1259 1260
      1261 1262 1263 1264
      1265 1266 1267 1268
      1269 1270 1271 1272
      1273 1274 1275 1276
      1277 1278 1279 1280
      1281 1282 1283 1284
      1285 1286 1287 1288
      1289 1290 1291 1292
      1293 1294 1295 1296
      1297 1298 1299 1300
      1301 1302 1303 1304
      1305 1306 1307 1308
      1309 1310 1311 1312
      1313 1314 1315 1316
      1317 1318 1319 1320
      1321 1322 1323 1324
      1325 1326 1327 1328
      1329 1330 1331 1332
      1333 1334 1335 1336
      1337 1338 1339 1340
      1341 1342 1343 1344
      1345 1346 1347 1348
      1349 1350 1351 1352
      1353 1354 1355 1356
      1357 1358 1359 1360
      1361 1362 1363 1364
      1365 1366 1367 1368
      1369 1370 1371 1372
      1373 1374 1375 1376
      1377 1378 1379 1380
      1381 1382 1383 1384
      1385 1386 1387 1388
      1389 1390 1391 1392
      1393 1394 1395 1396
      1397 1398 1399 1400
      1401 1402 1403 1404
      1405 1406 1407 1408
      1409 1410 1411 1412
      1413 1414 1415 1416
      1417 1418 1419 1420
      1421 1422 1423 1424
      1425 1426 1427 1428
      1429 1430 1431 1432
      1433 1434 1435 1436
      1437 1438 1439 1440
      1441 1442 1443 1444
      1445 1446 1447 1448
      1449 1450 1451 1452
      1453 1454 1455 1456
      1457 1458 1459 1460
      1461 1462 1463 1464
      1465 1466 1467 1468
      1469 1470 1471 1472
      1473 1474 1475 1476
      1477 1478 1479 1480
      1481 1482 1483 1484
      1485 1486 1487 1488
      1489 1490 1491 1492
      1493 1494 1495 1496
      1497 1498 1499 1500
      1501 1502 1503 1504
      1505 1506 1507 1508
      1509 1510 1511 1512
      1513 1514 1515 1516
      1517 1518 1519 1520
      1521 1522 1523 1524
      1525 1526 1527 1528
      1529 1530 1531 1532
      1533 1534 1535 1536
      1537 1538 1539 1540
      1541 1542 1543 1544
      1545 1546 1547 1548
      1549 1550 1551 1552
      1553 1554 1555 1556
      1557 1558 1559 1560
      1561 1562 1563 1564
      1565 1566 1567 1568
      1569 1570 1571 1572
      1573 1574 1575 1576
      1577 1578 1579 1580
      1581 1582 1583 1584
      1585 1586 1587 1588
      1589 1590 1591 1592
      1593 1594 1595 1596
      1597 1598 1599 1600
      1601 1602 1603 1604
      1605 1606 1607 1608
      1609 1610 1611 1612
      1613 1614 1615 1616
      1617 1618 1619 1620
      1621 1622 1623 1624
      1625 1626 1627 1628
      1629 1630 1631 1632
      1633 1634 1635 1636
      1637 1638 1639 1640
      1641 1642 1643 1644
      1645 1646 1647 1648
      1649 1650 1651 1652
      1653 1654 1655 1656
      1657 1658 1659 1660
      1661 1662 1663 1664
      1665 1666 1667 1668
      1669 1670 1671 1672
      1673 1674 1675 1676
      1677 1678 1679 1680
      1681 1682 1683 1684
      1685 1686 1687 1688
      1689 1690 1691 1692
      1693 1694 1695 1696
      1697 1698 1699 1700
      1701 1702 1703 1704
      1705 1706 1707 1708
      1709 1710 1711 1712
      1713 1714 1715 1716
      1717 1718 1719 1720
      1721 1722 1723 1724
      1725 1726 1727 1728
      1729 1730 1731 1732
      1733 1734 1735 1736
      1737 1738 1739 1740
      1741 1742 1743 1744
      1745 1746 1747 1748
      1749 1750 1751 1752
      1753 1754 1755 1756
      1757 1758 1759 1760
      1761 1762 1763 1764
      1765 1766 1767 1768
      1769 1770 1771 1772
      1773 1774 1775 1776
      1777 1778 1779 1780
      1781 1782 1783 1784
      1785 1786 1787 1788
      1789 1790 1791 1792
      1793 1794 1795 1796
      1797 1798 1799 1800
      1801 1802 1803 1804
      1805 1806 1807 1808
      1809 1810 1811 1812
      1813 1814 1815 1816
      1817 1818 1819 1820
      1821 1822 1823 1824
      1825 1826 1827 1828
      1829 1830 1831 1832
      1833 1834 1835 1836
      1837 1838 1839 1840
      1841 1842 1843 1844
      1845 1846 1847 1848
      1849 1850 1851 1852
      1853 1854 1855 1856
      1857 1858 1859 1860
      1861 1862 1863 1864
      1865 1866 1867 1868
      1869 1870 1871 1872
      1873 1874 1875 1876
      1877 1878 1879 1880
      1881 1882 1883 1884
      1885 1886 1887 1888
      1889 1890 1891 1892
      1893 1894 1895 1896
      1897 1898 1899 1900
      1901 1902 1903 1904
      1905 1906 1907 1908
      1909 1910 1911 1912
      1913 1914 1915 1916
      1917 1918 1919 1920
      1921 1922 1923 1924
      1925 1926 1927 1928
      1929 1930 1931 1932
      1933 1934 1935 1936
      1937 1938 1939 1940
      1941 1942 1943 1944
      1945 1946 1947 1948
      1949 1950 1951 1952
      1953 1954 1955 1956
      1957 1958 1959 1960
      1961 1962 1963 1964
      1965 1966 1967 1968
      1969 1970 1971 1972
      1973 1974 1975 1976
      1977 1978 1979 1980
      1981 1982 1983 1984
      1985 1986 1987 1988
      1989 1990 1991 1992
      1993 1994 1995 1996
      1997 1998 1999 2000
      2001 2002 2003 2004
      2005 2006 2007 2008
      2009 2010 2011 2012
      2013 2014 2015 2016
      2017 2018 2019 2020
      2021 2022 2023 2024
      2025 2026 2027 2028
      2029 2030 2031 2032
      2033 2034 2035 2036
      2037 2038 2039 2040
      2041 2042 2043 2044
      2045 2046 2047 2048
      2049 2050 2051 2052
      2053 2054 2055 2056
      2057 2058 2059 2060
      2061 2062 2063 2064
      2065 2066 2067 2068
      2069 2070 2071 2072
      2073 2074 2075 2076
      2077 2078 2079 2080
      2081 2082 2083 2084
      2085 2086 2087 2088
      2089 2090 2091 2092
      2093 2094 2095 2096
      2097 2098 2099 2100
      2101 2102 2103 2104
      2105 2106 2107 2108
      2109 2110 2111 2112
      2113 2114 2115 2116
      2117 2118 2119 2120
      2121 2122 2123 2124
      2125 2126 2127 2128
      2129 2130 2131 2132
      2133 2134 2135 2136
      2137 2138 2139 2140
      2141 2142 2143 2144
      2145 2146 2147 2148
      2149 2150 2151 2152
      2153 2154 2155 2156
      2157 2158 2159 2160
      2161 2162 2163 2164
      2165 2166 2167 2168
      2169 2170 2171 2172
      2173 2174 2175 2176
      2177 2178 2179 2180
      2181 2182 2183 2184
      2185 2186 2187 2188
      2189 2190 2191 2192
      2193 2194 2195 2196
      2197 2198 2199 2200
      2201 2202 2203 2204
      2205 2206 2207 2208
      2209 2210 2211 2212
      2213 2214 2215 2216
      2217 2218 2219 2220
      2221 2222 2223 2224
      2225 2226 2227 2228
      2229 2230 2231 2232
      2233 2234 2235 2236
      2237 2238 2239 2240
      2241 2242 2243 2244
      2245 2246 2247 2248
      2249 2250 2251 2252
      2253 2254 2255 2256
      2257 2258 2259 226
```

## Drawback to Profiles

- Profiles only as representative as the variation in the training sets.
- Thus, there is a bias in the profile towards the training data.
- Training sets can be erroneous if not carefully constructed

7

## Calculating Profiles

- Each cell is the **log-odds** score
  - The value of an individual cell is calculated as the log odds score of finding a particular residue in a particular location in an alignment divided by the probability of aligning the two amino acids by random chance using a particular scoring scheme (such as PAM250, BLOSUM80, ...).
    - PAM (Percent Accepted Mutation)
    - BLOSUM (Blocks Substitution Matrix)
  - Additional penalties must be calculated for gap opening and gap extension in the profile as well.
- Some methods take in sequence weights as well
  - One method (average method) weighs the proportion of the amino acids found in a particular column, and weights the score of matching the consensus residue at a given position to that particular residue.

8

## Shannon Entropy

- One method to calculate the observed column variation given the expected variation in the evolutionary model is to use an information measure known as **entropy**.
  - Entropy is the amount of information of the observed column variation if expected variation in the evolutionary model is known
- The smaller the entropy, the more conserved a column is.

9

## Entropy...

- The entropy (**H**) for a single column is calculated by the following formula:

$$H = - \sum_{\text{residues}(a)} f_a \log(p_a)$$

- **a**: is a residue (amino acid),
- **f<sub>a</sub>**: frequency of residue **a** in a column,
- **p<sub>a</sub>**: probability (expected frequency) of residue **a** in that column

10

## ...Entropy...

- **H** is calculated for each 20 ancestor amino acids and for a large number of evolutionary distances (PAM1, PAM2, PAM4, ...).
- The distance that gives the minimum value for **H** for each column-possible ancestor combination is the best estimate of the distance that generates the column diversity from that ancestor.
- This analysis provides 20 possible models (**M<sub>a</sub>** for **a** = 1,2,3,...,20) as to how the amino acid frequencies in a column (**F**) may have originated.

11

## ...Entropy...

- The next step in the evolutionary profile construction determines the extent to which each **M<sub>a</sub>** predicts **F** by the Bayes conditional probability analysis.

$$P(M_a|F) = P(M_a) \times P(F|M_a) / \sum_{\text{all } a's} P(M_a) \times P(F|M_a)$$

- where the prior distribution **P(M<sub>a</sub>)** is given by the background amino acid frequencies and

$$P(F|M_a) = P_{aa1}^{faa1} \times P_{aa2}^{faa2} \times P_{aa3}^{faa3} \dots P_{aa20}^{faa20}$$

- i.e., the product of the expected amino acid frequencies in **M<sub>a</sub>** raised to the power of the fraction observed for each amino acid in the msa column.

12

## ...Entropy

- From  $P(M_a|F)$ , the weights for each of the 20 possible distributions that give rise to the msa column diversity are calculated as follows:

$$W_a = P(M_a|F) - P(M_{random}|F)$$

- where  $W_a$  is the weight given to  $M_a$  and  $P(M_{random}|F)$  is calculated as above using amino acid distribution.

13

## Log-odds score...

- Another measure of creating a profile is by using **log-odds score**.
- In this method,
  - the  $\log_2$  of the ratio of observed/background frequencies is calculated for each position.
  - What results is the amount of information available in an alignment given in bits.
- A new sequence can then be searched to see if it possibly contains the motif.
- Profiles can also indicate log-odds score:
  - $\text{Log}_2(\text{observed} \div \text{expected})$
- Result is a bit score

14

## ...Log-odds score

- The **log odds scores** for the profile (**Profile<sub>ij</sub>**) are given by

$$\text{Profile}_{ij} = \log \left[ \sum_{\text{all } a's} (W_{ai} \times P_{aij}) / P_{\text{random}j} \right]$$

where

- $W_{ai}$  is the weight of an ancestral amino acid  $a$  at row  $i$  in the profile,
- $P_{aij}$  is the frequency of amino acid  $j$  in the PAM amino acid distribution that best matches at row  $i$ ,
- $P_{\text{random}j}$  is the background frequency of amino acid  $j$ .

15

## BLOCK Analysis...

- Blocks are similar to profiles in the sense that
  - they represent locally conserved regions within a MSA.
- However, the difference is that ...
  - blocks lack insert and delete (**indels**) positions in the sequences.
  - Instead, every column includes only matches and mismatches
- Blocks can be determined either
  - by performing a multiple sequence alignment, or
  - by searching a database for similar sequences of the same length.

16

## ...BLOCK Analysis...

- Generally determined by performing multiple alignment first
- Ungapped regions are then separated into blocks
- Algorithms have been developed for searching for blocks

17

## ...BLOCK Analysis

- Statistical approaches to finding the most alike sequences have been proposed, such as
  - the Expectation-Maximization algorithms and
  - the Gibbs sampler.
- In any case, once a set of blocks has been determined, the information contained within the block alignment can be displayed as a sequence profile.

18

## BLOCKS Programs

- A global sequence alignment will usually contain ungapped regions that are aligned between multiple sequences.
- These regions can be extracted to produce blocks.
- Two widely used programs:
  - BLOCKS
  - eMOTIF

[http://www.blocks.fhcrc.org/blocks/process\\_blocks.html](http://www.blocks.fhcrc.org/blocks/process_blocks.html)

<http://dna.stanford.edu/emotif/>

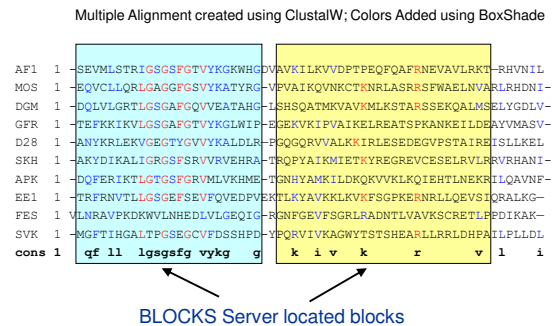
## Example...

- 10 Truncated Kinase proteins
  - Approximately 75 residues in length
    - A protein kinase is a kinase enzyme that modifies other proteins by chemically adding phosphate groups to them (phosphorylation).
    - The human genome contains about 500 protein kinase genes and they constitute about 2% of all human genes.
    - Protein kinases are also found in bacteria and plants.
    - Up to 30% of all human proteins may be modified by kinase activity, and kinases are known to regulate the majority of cellular pathways, especially those involved in signal transduction.

## ...Example...

```
>D28      CD28  S. CEREVISIAE CELL CYCLE CONTROL PROTEIN KINASE
ANYKRLKVGEGTYGVVYKALDLRPQQQGVVVALKKIRLESEDEGVFPSTAIRESLKL
>SKH      SKH  HELA MYSTERY PUTATIVE PROTEIN KINASE
AKYDIKALIGRGSFRRVVRVEHRAITRQPYAIKMIETKYREGREVCESELRLVLRVRHANI
>APK      CAPK BOVINE CARDIAC MUSCLE CYCLIC AMP-DEPENDENT (ALPHA)
DQFERIKTLGTGSFGRVMLVKHMETGNHYAMKILDKQVVKLKQIEHTLNEKRILQAVNF
>EE1      WEE1  S. POMBE MITOTIC INHIBITOR
TRFRNVTLLSGSEFSEVFQVEDPVKTLKAVKVLKLVKFSFGPKERNRLQEVSIQRLKG
>GFR      EGFR HUMAN EPIDERMAL GROWTH FACTOR RECEPTOR
TEFKKIKVLGSGAFGTIVYKGLWIPGEKVKIPVAIKELREATSPKANKEILDEAYVMASV
>DGM      PDGF RECEPTOR, MOUSE KINASE REGION
DQLVLGRTLGSAGFQGVVEATAHGLSHQATMKVAVMLKSTARSSEKQALMSLYGDLV
>FES      THIS IS VFES TYROSINE KINASE
VLNRAVPKDKWVLNHDVLVLGEQIGRNFGEVFSGRRLADNTLVAVKSCRETLPDIKAK
>AF1      RAF1 HUMAN C-RAF-1 ONCOGENE
SEVMLSTRIGSGSGFTVYKQVVAIKQVKNCTKNRLASRRSFVAELNVARLRHDNI
>MOS      CMOS HUMAN C-MOS ONCOGENE
EQVCLLQRLGAGGFGSVYKATYRGVFAIKQVKNCTKNRLASRRSFVAELNVARLRHDNI
>SVK      HSVK HERPES SIMPLEX VIRUS PUTATIVE PROTEIN KINASE
MGFTIHGALTPGSEGCVFDSHPDYPQVRVIVKAGWYTTSTSEARLLRLDHPAILPLLDL
```

## ...Example...



## ...Example...

```
* Taking this alignment, blocks can be generated using
the BLOCKS server:

ID      x6676xbli; BLOCK
AC      x6676xbliA; distance from previous blocks=(1,1)
DE      ../tmp/6676.blin
BL      UNK motif; width=24; seqs=10; 99.5%=0; strength=0
AF1      ( 1) SEVMLSTRIGSGSGFTVYKQVVAIKQVKNCTKNRLASRRSFVAELNVARLRHDNI 41
MOS      ( 1) EQVCLLQRLGAGGFGSVYKATYRGVFAIKQVKNCTKNRLASRRSFVAELNVARLRHDNI 48
DGM      ( 1) DQLVLGRTLGSAGFQGVVEATAHGLSHQATMKVAVMLKSTARSSEKQALMSLYGDLV 49
GFR      ( 1) TEFKKIKVLGSGAFGTIVYKGLWIPGEKVKIPVAIKELREATSPKANKEILDEAYVMASV 41
D28      ( 1) ANYKRLKVGEGTYGVVYKALDLRPQQQGVVVALKKIRLESEDEGVFPSTAIRESLKL 61
SKH      ( 1) AKYDIKALIGRGSFRRVVRVEHRAITRQPYAIKMIETKYREGREVCESELRLVLRVRHANI 54
APK      ( 1) DQFERIKTLGTGSFGRVMLVKHMETGNHYAMKILDKQVVKLKQIEHTLNEKRILQAVNF 46
EE1      ( 1) TRFRNVTLLSGSEFSEVFQVEDPVKTLKAVKVLKLVKFSFGPKERNRLQEVSIQRLKG 55
FES      ( 1) LNRRAVPKDKWVLNHDVLVLGEQIGRNFGEVFSGRRLADNTLVAVKSCRETLPDIKAK 100
SVK      ( 1) MGFTIHGALTPGSEGCVFDSHPDYPQVRVIVKAGWYTTSTSEARLLRLDHPAILPLLDL 73
//
```

## ...Example

```
ID      x6676xbli; BLOCK
AC      x6676xbliB; distance from previous blocks=(2,2)
DE      ../tmp/6676.blin
BL      UNK motif; width=28; seqs=10; 99.5%=0; strength=0
AF1      ( 27) AVKILKVVDPTEPQFQAFRNEVAVLRKT 87
MOS      ( 27) PVAIKQVKNCTKNRLASRRSFVAELNVA 75
DGM      ( 27) SHSQATMKVAVMLKSTARSSEKQALMS 92
GFR      ( 27) GEKVKIPVAIKELREATSPKANKEILDE 83
D28      ( 27) PQGQQRVVALKKIRLESEDEGVFPSTAIR 83
SKH      ( 27) RQPYAIKMIETKYREGREVCESELRLVLR 74
APK      ( 27) GNHYAMKILDKQVVKLKQIEHTLNEKR 85
EE1      ( 27) TLKYAVKVLKLVKFSFGPKERNRLQEVSI 77
FES      ( 27) GNFGEVFSGRRLADNTLVAVKSCRETLP 100
SVK      ( 27) PQRVIVKAGWYTTSTSEARLLRLDHPA 92
//
```

## Statistical Methods for Aiding Alignments

- Commonly used methods for locating motifs:

- Expectation-Maximization (EM)

- Gibbs Sampling

## Expectation-Maximization...

- EM algorithm has been used to identify both conserved domains in unaligned proteins and protein-binding sites in unaligned DNA sequences, including sites that may include gaps
- In the EM algorithms,
  - the starting point is a set of sequences expected to have a common sequence pattern that may not be easily detectible.
  - An initial guess is made as to the location and size of the site of interest in each of the sequences.
  - These initial sites are then aligned.
  - Approximate length of signal must be given
- Randomly assign locations of this motif in each sequence

25

26

## ...Expectation-Maximization...

- The EM algorithm consists of two steps, which are repeated consecutively:
  - Expectation Step
    - In the expectation step, background residue frequencies are calculated based on those residues that are not in the initially aligned sites.
    - Column specific residues are calculated for each position in the initial motif alignment.
    - Using this information, the probability of finding the site at any position in the sequences can then be calculated.
    - Residues not in a motif are background
  - Frequencies used to determine probability of finding site at any position in a sequence to fit motif model

27

## ...Expectation-Maximization

- Maximization Step

- In the maximization step, the counts of residues for each position in the site as found in the expectation step are used to calculate the location within each sequence that maximally aligns to the motif pattern calculated in the expectation step.
- This is done for each of the sequences.
- Once a new motif location has been calculated, the expectation step is repeated.
- This cycle continues until the solution converges.

28

### Example of EM - initial alignment...

[illegible]

begin with an  
initial, random  
alignment:

### ...Example of EM - Residue Counts...

- From this alignment, the frequency of each base occurring is calculated.
- In this case, the motif we are searching for is six bases wide.
  - Therefore, we need to calculate seven different sets of frequencies:
    - One for the background,
    - one for each of the columns in the motif.
- Calculating the total counts, we get:

Nucleotide	Motif Position (0 = Background)							
e	0	1	2	3	4	5	6	Total
A	279	6	12	6	6	11	7	48
C	280	8	3	5	7	7	7	37
G	225	9	8	10	7	5	8	47
T	262	6	6	8	9	6	7	42
Total	1046	29	29	29	29	29	29	174

29

30

### ...Example of EM - Residue Frequencies...

- After calculating the observed counts for each of the positions, we can convert these to observed frequencies:

Nucleotide	Motif Position (0 = Background)						
	0	1	2	3	4	5	6
A	0.267	0.209	0.414	0.209	0.209	0.379	0.241
C	0.267	0.276	0.103	0.172	0.241	0.241	0.241
G	0.216	0.310	0.276	0.345	0.241	0.172	0.276
T	0.250	0.209	0.209	0.276	0.310	0.209	0.241

– Frequency of nucleotide *a* for the background (Col0):

# of nucleotide *a* in Col0 in Row*a* / # of all nucleotides in Col0

– Frequency of *a* in Col*c*:

# of nucleotide *a* in Col*c* / # of all nucleotides in Col*c*

31

### ...Example of EM - Residue Frequencies...

- However, in order to alleviate the issue of **zero counts** and overtraining of the data, **pseudocounts** are introduced to the observed counts:

– In this case, frequency of nucleotide *a* in Col*c*:

$$P_{ca} = (n_{ca} + b_{ca}) / (N_c + B_c)$$

$P_{ca}$ : Probability of residue *a* in column *c*;  $n_{ca}$ : count of *a*'s in column *c*;  $b_{ca}$ : pseudocount of *a*'s in column *c*;

$N_c$ : total count in column *c*;  $B_c$ : total pseudocount in column *c*

– Choosing a pseudocount is arbitrary

– For example, assuming that 4 nucleotides have equal probabilities, if total pseudocount ( $B_c$ ) is chosen as 1, pseudocount of each nucleotide will be  $b_{ca} = B_c/4$ .

– Note that a different pseudocount scheme is used in the following table

Nucleotide	Motif Position (0 = Background)						
	0	1	2	3	4	5	6
A	0.267	0.256	0.296	0.256	0.256	0.289	0.263
C	0.267	0.263	0.230	0.243	0.256	0.256	0.256
G	0.216	0.240	0.233	0.246	0.226	0.213	0.233
T	0.250	0.241	0.241	0.254	0.261	0.241	0.248

32

### ...Example of EM - Maximization Step...

- In the expectation step, the residue frequencies for the motif are used to estimate the composition of the motif site.
- The expectation step attempts to maximally discriminate between sequence within and not within the site.
- For each sequence, each possible motif location is considered in order to find the most probable location given the current motif.
- Consider the first sequence:
  - TCAGAACCCAGTTATATATTCATTTCCTTCTCCACTCCT
  - There are 41 residues; 41 - 6 + 1 = 36 sites to consider
- Starting from the first site (TCAGAA), 36 scores for the first sequence are calculated

33

### ...Example of EM - Residue Frequencies...

- Let us consider the eighth site CAGTTA.

– In order to calculate site score, observed frequency table is used:

Nucleotide	Motif Position (0 = Background)						
	0	1	2	3	4	5	6
A	0.267	0.256	0.296	0.256	0.256	0.289	0.263
C	0.267	0.263	0.230	0.243	0.256	0.256	0.256
G	0.216	0.240	0.233	0.246	0.226	0.213	0.233
T	0.250	0.241	0.241	0.254	0.261	0.241	0.248

- Position:

1

2

3

4

5

6

C

A

G

T

T

A

$$S_{\text{CAGTTA}} = 0.263 \times 0.296 \times 0.246 \times 0.261 \times 0.241 \times 0.263$$

$$S_{\text{CAGTTA}} = 0.000317$$

34

### ...Example of EM - Maximization Step...

	1	2	3	4	5	6	1 <sup>2</sup> 2 <sup>3</sup> 3 <sup>4</sup> 4 <sup>5</sup> 5 <sup>6</sup>	RANDOM	ODDS
TCAGAA	241	230	256	226	289	263	0.000244	0.000274	0.89
CAGAAC	263	296	246	256	289	256	0.000163	0.000362	1.00
AGAACC	256	233	256	256	256	256	0.000256	0.000362	0.71
GAACCA	240	296	256	256	256	263	0.000113	0.000362	0.87
AACCAG	256	296	243	256	289	233	0.000117	0.000362	0.88
ACCAGT	256	230	243	256	213	348	0.000193	0.000274	0.71
CCAGTT	263	230	256	226	241	348	0.000209	0.000257	0.81
CAGTTA	263	296	246	261	241	263	0.000317	0.000257	1.23
AGTTAT	256	233	254	263	289	248	0.000283	0.000241	1.18
GTATAT	240	241	254	256	241	263	0.000238	0.000241	0.99
TTATAA	241	241	256	261	289	263	0.000295	0.000297	0.99
TATAAA	241	296	254	256	289	263	0.000353	0.000297	1.19
ATAAAT	256	241	256	256	289	248	0.000280	0.000318	0.91
TAAATT	241	296	256	256	241	248	0.000279	0.000297	0.94
AAATTT	256	296	256	261	241	248	0.000303	0.000297	1.02
AAATTA	256	296	254	261	241	263	0.000318	0.000297	1.07
ATTAT	256	241	254	263	289	248	0.000293	0.000278	1.05
TTATTC	241	241	254	256	241	256	0.000233	0.000278	0.84

35

### ...Example of EM - Maximization Step...

TTATCA	241	241	256	261	256	263	0.000261	0.000297	0.88
TATCAT	241	296	254	256	289	248	0.000332	0.000297	1.12
ATCAT	256	241	243	256	241	248	0.000229	0.000297	0.77
TCATTT	241	230	256	261	241	248	0.000221	0.000278	0.80
CATTTC	263	296	254	261	241	256	0.000318	0.000297	1.07
ATTTC	256	241	254	261	256	256	0.000268	0.000297	0.90
TTTCCT	241	241	254	256	256	248	0.000240	0.000278	0.86
TTCTTT	241	241	243	256	241	248	0.000216	0.000278	0.78
TCCTTC	241	230	243	261	241	256	0.000217	0.000297	0.73
CCTTCT	263	230	254	261	256	248	0.000255	0.000297	0.86
CTTCTC	263	241	254	256	241	256	0.000254	0.000297	0.86
TTCTCC	241	241	243	261	256	256	0.000241	0.000297	0.81
TTCTCA	241	230	254	256	256	263	0.000243	0.000318	0.76
CTCCAC	263	241	243	256	289	256	0.000292	0.000339	0.86
TCCACT	241	230	243	256	256	248	0.000219	0.000318	0.69
CCACTC	263	230	256	256	241	256	0.000245	0.000339	0.72
CACCTC	263	296	243	261	256	248	0.000324	0.000339	0.95
ACTCCT	256	230	254	256	256	248	0.000243	0.000318	0.76

36

## ...Example of EM - Maximization Step...

- The six base site **CAGTTA** beginning at base 8 is calculated to have the highest odds probability.
- Therefore, it is chosen as the new site in sequence 1.
- This is repeated for each of the sequences.
- In the maximization step, the newly chosen sites for each of the sequences are used to recalculate the frequency table.
- The expectation/maximization cycle is then repeated, until the results converge on a set of motifs.

37

## ...Example of EM - Maximization Step

- Before:
  - Random Alignment
- TCAGAACCAAGTTATAA**ATTAT**CATTTCCTTCTCCACTCCT
- After:
  - Maximal location (given random motif alignment) (first round)
- TCAGAAC**CAGTTA**TAA**ATTAT**CATTTCCTTCTCCACTCCT

38

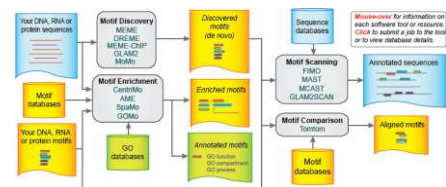
## Available E-M Programs

- MEME – Uses E-M algorithms as explained
  - Multiple EM for Motif Elicitation (MEME) is a program developed that uses the expectation-maximization methods as described previously.
    - ParaMEME searches for blocks using the EM algorithm,
    - MetaMEME searches for profiles using Hidden Markov Models (HMMs).
- MEME locates one or more ungapped patterns in a single DNA or protein sequence, or in a series of sequences.
- A search is conducted on a variety of motif widths in order to determine the most likely width for the profile.
  - This likelihood is based on the log likelihood score calculated after the EM algorithm.

39

## The MEME Suite

- Motif-based sequence analysis tools



- <http://meme-suite.org/index.html>

40

## MEME Software

- One of three types of motif models can be chosen:
  - OOPS (One expected Occurrence Per Sequence)
    - simplest model type since it assumes that there is exactly one occurrence per sequence of the motif in the dataset.
  - ZOOPS (Zero or One expected Occurrence Per Sequence)
    - generalization of OOPS
    - assumes zero or one motif occurrences per dataset sequence
  - TCM (Two-Component Mixture)
    - assumes that there are zero or more non-overlapping occurrences of the motif in each sequence in the dataset

– Bailey, Timothy L. and Charles Elkan, "The Value of Prior Knowledge in Discovering Motifs with MEME." Proceedings. International Conference on Intelligent Systems for Molecular Biology 3 (1995): 21-9.

– [https://tbailey.bitbucket.io/papers/cs95\\_143.pdf](https://tbailey.bitbucket.io/papers/cs95_143.pdf)

41

## MEME Software

- Various prior knowledge can be added to MEME, including
  - the expected number of motifs,
  - the expected length of the motif,
  - whether or not the motif is palindromic
    - only applicable for DNA sequences

42

## Gibbs Sampling...

- Similar in nature to the EM algorithms.
  - Combines both EM and simulated annealing techniques in order to determine a maximal local alignment of multiple sequences.
  - Goal is to find most probable pattern by sampling from motif probabilities to maximize model÷background probabilities
    - The idea behind Gibbs sampling is to determine the most probable pattern common to all of the sequences by sliding them back and forth until the ratio of the motif probability to the background probability is a maximum.

43

## ...Gibbs Sampling...

- Predictive Update Step
  - Random motif start position chosen for all sequences except one
  - Initial alignment used to calculate residue frequencies for motif and background
  - Similar to the Expectation Step of EM
- Sampling Step
  - Model probability÷background probability normalized and weighted
  - Motif start position chosen based on a random sampling with the given weights
  - Different than EM algorithm

44

## ...Gibbs Sampling

- Process repeated until residue frequencies in each column do not change
- The sampling step is then repeated for a different initial random alignment
  - Sampling allows escape from local maxima
  - Employs a shifting routine that will take a current multiple motif alignment, and shift it a few bases to the left or the right, in order to see if only part of the motif is being found
  - A range of motif sizes can be explored in Gibbs sampling as well
- Gibbs sampling can be extended
  - to search for multiple motifs in the same set of sequences,
  - to find a pattern in only a fraction of the sequences.
- In addition, certain model-specific parameters can be enforced, such as palindromic sequences

45

## Hidden Markov Models...

- Hidden Markov Models (HMMs)
  - probabilistic models for studying sequences of symbols.
- HMMs can model matches, mismatches, insertions and deletions of symbols.
- HMMs have been deeply rooted in speech recognition problems.
- In speech recognition, the problem is the phonemes (or words) that have been spoken in a particular time frame.

46

## ...Hidden Markov Models...

- Consider the difficulty.
  - Everyone you meet has a different voice.
  - Everyone speaks with a slight variation
    - this might be caused by an accent, the person having a cold, or differences in physiological development.
- However, humans are able to distinguish what the speaker is saying.
  - The idea behind speech recognition is to take in a spoken word and to try to fit it to a specific model of possible words.
    - This may in fact be close to what the brain does

47

## ...Hidden Markov Models

- Problems in sequence analysis are similar.
- For instance,
- given an amino acid sequence, we may want to determine the protein family to which it belongs.
- The amino acid sequence can be treated similarly to the speech signal in a given frame, and the amino acids can be treated as the phonemes.

48

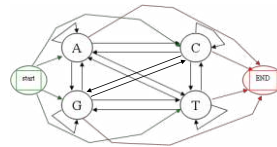


## Markov Chain

- A probabilistic model that generates a sequence where the probability of a symbol depends upon the previous symbol.
  - A traffic light is an example of a Markov chain.
- A Markov Chain can be used to model a random DNA sequence, where there are four states:
  - A, C, G, T
    - one for each letter in the alphabet.
- When we are given a certain state, there is a transition from that state to another state with an associated probability
  - called a **transition probability**.
- An example Markov Chain can be drawn as follows:

49

## Markov Chain



- The key property of a Markov chain is that
  - the probability of a symbol  $S$  at position  $p(S_p)$  depends only upon the previous symbol  $S$  at position  $p-1(S_{p-1})$ , and not on the entire previous sequence.
- Since the probability of a symbol is dependent upon the previous symbol, a prime example for the use of **Markov chains** is in the detection of **CpG islands**, which are rich in the dinucleotide **CG**.
  - **CpG (CG) islands** is a short stretch of DNA in which the frequency of the CG sequence is higher than other regions.
    - "p" simply indicates that "C" and "G" are connected by a phosphodiester bond.

50

## Markov Chain

- The process of methylation in biological systems will typically convert the nucleotide **C** to a **T** with a high probability when a **CG** nucleotide is encountered.
  - As a result, there will be an overabundance of the dinucleotide **TG**, and an underabundance of the dinucleotide **CG**.
- If we ignore the start and end states for now, we can see that there are sixteen different transitions.
  - A study of regions of genomic DNA has determined normal genomic transition probabilities to be the following.
    - where the **FROM** node is labeled along the rows to the left, and the **TO** node is labeled along the columns above:

51

## Markov Chain

	A	C	G	T
A	0.300	0.205	0.285	0.210
C	0.322	0.298	0.078	0.302
G	0.248	0.246	0.298	0.208
T	0.177	0.239	0.292	0.292

- The model shown above can then assign these weights to the edges of the graph

52

## Markov Chain

- In some regions of the genome, such as the promoter region of genes, methylation is suppressed.
  - In these regions, the dinucleotide **CG** is found in greater quantities.
- In fact, the nucleotides **C** and **G** are found to a greater degree than elsewhere in the genome.
  - A study of regions of genomic DNA where **CpG islands** exist has determined the transition probabilities to be the following:

53

## Markov Chain

	A	C	G	T
A	0.180	0.274	0.426	0.120
C	0.171	0.368	0.274	0.188
G	0.161	0.339	0.375	0.125
T	0.079	0.355	0.384	0.182

- A new model just like the one above can have its transition properties assigned according to the new table.
- Now we have two different models:
  - the first where **CpG islands** are absent,
  - the second where **CpG islands** are present.

54

## Markov Chain

- Let's call the first model the **non-CpG model** and the second model the **CpG model**.
- Given a new sequence, how would we determine whether it belongs to the **non-CpG model** or the **CpG model**?
- Remember, the key property of a Markov chain
  - the probability of a symbol **S** at position  $p(S_p)$  depends only upon the previous symbol **S** at position  $p-1(S_{p-1})$ .
    - not on the entire previous sequence.

55

## Markov Chain

- Therefore, to find the probability that a sequence fits a model,
  - you would multiply all of the conditional probabilities:
$$P(x) = P(x_L|x_{L-1})P(x_{L-1}|x_{L-2}) \dots P(x_2|x_1)P(x_1)$$
- which can be rewritten as:

$$P(x) = P(x_1) \prod_{i=2}^L a_{x_{i-1}x_i}$$

- where  $a_{x_{i-1}x_i}$  is the probability from residue at position  $i-1$  to the residue at position  $i$

56

## Markov Chain

- Let's consider for now that in the **non-CpG model**,  $P(A) = P(T) = 0.3$ ;  $P(C) = P(G) = 0.2$ ,
  - so that **A** and **T** are more probable.
- In the **CpG model**, consider  $P(A) = P(C) = P(G) = P(T) = 0.25$ .
- Now consider the sequence: **GGCGACG**
- The probability for this sequence:
 
$$P(G)P(G|G)P(C|G)P(G|C)P(A|G)P(C|A)P(G|C)$$

57

## Markov Chain

- For the **non-CpG model** can be calculated as:
 
$$(0.20)(0.298)(0.246)(0.078)(0.248)(0.205)(0.078) = 0.000000453499$$
- For the **CpG model** can be calculated as:
 
$$(0.25)(0.375)(0.339)(0.274)(0.161)(0.274)(0.274)(0.125) = 0.0010526$$
- Given this information, it is more likely that this sequence fits the **CpG model**.
- One thing to note is how quickly the probability gets to zero.
  - This shows the importance of using log statistics.

58

## Using Markov models for discrimination

- How different the **non-CpG** and **CpG models** are in relation to each other?
  - If they are not different enough, then there is not enough information to determine from which model a particular sequence is derived.
- In order to test whether we are able to discriminate between the two models, a log ratio is taken for each of the scores in the two previous tables to create a third table, where each entry,  $x$ , in the new table is equal to:
 
$$\log_2(P(x|\text{CpG model}) / P(x|\text{non-CpG model}))$$

59

## Using Markov models for discrimination

- The resulting table is as follows:

	A	C	G	T
A	-0.740	0.419	0.580	-0.803
C	-0.913	0.302	1.812	-0.685
G	-0.624	0.461	0.331	-0.730
T	-1.169	0.573	0.393	-0.679

- Using this log-odds ratio table as the scores, we can then see that
  - a sequence with a negative score will belong to the **non-CpG model**,
  - a sequence with a positive score will belong to the **CpG model**.

60

## Position Specific Scoring Matrix (PSSM)

- **Position Specific Scoring Matrices** incorporate information theory in order to gain a measure of how much information is contained within each column of a multiple alignment.
  - The information contained within a **PSSM** is a logarithmic transformation of the frequency of each residue in the motif.
- One problem with creating a model of a sequence alignment that is then used to search databases is that there is a bias towards the training data
  - Some residues may be underrepresented
  - Other columns may be too conserved

61

## Pseudocounts...

- Solution:
  - Introduce **Pseudocounts** to get a better indication
- The goal of adding **pseudocounts** is to obtain an improved estimate of the probability  $p_{ca}$  that amino acid  $a$  is in column  $c$  in all occurrences of the blocks, and not just the ones in the present sample.
- The current estimate of  $p_{ca}$  is  $f_{ca}$ , the frequency of counts in the data.
- A simplified Bayesian prediction improves the estimate of  $p_{ca}$  by adding prior information in the form of **pseudocounts**

62

## ...Pseudocounts

- Now the estimated probability is changed from a frequency of counts in the data to the following form:

$$P_{ca} = \frac{n_{ca} + b_{ca}}{N_c + B_c}$$

- $P_{ca}$ : Probability of residue  $a$  in column  $c$
- $n_{ca}$ : count of  $a$ 's in column  $c$
- $b_{ca}$ : pseudocount of  $a$ 's in column  $c$
- $N_c$ : total count in column  $c$
- $B_c$ : total pseudocount in column  $c$
- These probabilities are then converted into a **log-odds** form (usually  $\log_2$  so the information can be reported in bits) and placed in the PSSM .

63

## Searching PSSMs

- In order to search a sequence against a **PSSM**, the value for the first residue in the sequence occurring in the first column is calculated by searching the **PSSM**.
- Similarly, the value for the residue occurring in each column is calculated.
- These values are added (since they are logarithms) to produce a summed log odds score,  $S$ .
- This score can be converted to an odds score using the formula  $2^S$ .
- The odds scores for the motif beginning at each position can be summed together and normalized to produce a probability of the motif occurring at each location.

64

## Information in PSSMs

- Information theory can give an appreciation for the amount of information contained within each sequence.
- When there is no information contained within a column, the amount of uncertainty can be measured as
  - $\log_2 20 = 4.32$  for amino acids (20 amino acids)
  - $\log_2 4 = 2$  for nucleic acid sequences (4 nucleotides)
- If only one amino acid is found in a particular column, then the uncertainty is 0 (there is only one choice).
- If there are two amino acids occurring with equal probability, then there is an uncertainty to deciding which residue it is.

65

## Measure of Uncertainty

- The amount of uncertainty for a particular column is measured as the **entropy**, as introduced previously

$$H_c = - \sum_{\text{residues } (a)} f_{ac} \log(p_{ac})$$

- The uncertainty for the whole **PSSM** can be calculated as a sum over all columns:

$$H_c = \sum_{\text{all columns}} H_c$$

66

## Relative Entropy

- In addition to the entropy measure given before, a **relative entropy** measure could be calculated as well.
  - Relative entropy** takes into account not only the data in the columns of the motif, but also the overall composition of the organism being studied.
- Relative entropy** can be measured as:

$$R_C = - \sum_{\text{residue}(a)} f_{ac} \log_2(p_{ac}/b_a)$$

- $b_a$  is background frequency of residue  $a$  in the organism

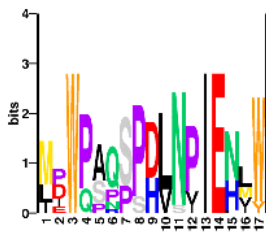
## Sequence Logos...

- One way to look at a particular **PSSM** is to view it visually.
  - Sequence logos** are one way to do so, by illustrating the information in each column of a motif.
- Such a graph can indicate which residues and which columns are the most important as far as sequence conservation is concerned.
  - The height of the logo is calculated as the amount by which uncertainty has been decreased
  - If the frequency in the column is less than the frequency in the background, then a negative relative entropy can be computed, which can be shown by an inverted character in the logo.

67

68

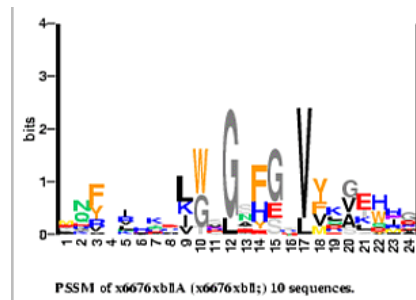
## ...Sequence Logos...



Logo of Gibbs Block D (Tcl) 9 sequences

69

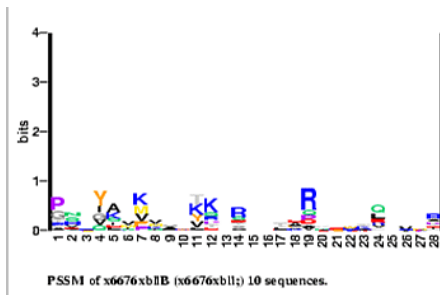
## ...Sequence Logos...



PSSM of x6676xb1A (x6676xb1z) 10 sequences.

70

## ...Sequence Logos



PSSM of x6676xb1B (x6676xb1z) 10 sequences.

71

## Sequence Editors

- Allow manual editing of alignments
- Add color to alignments
- Prepare images for publication
- Some sequence editors:
  - BoxShade [http://www.ch.embnet.org/software/BOX\\_form.html](http://www.ch.embnet.org/software/BOX_form.html)
  - Serial Cloner [http://serialbasics.free.fr/Serial\\_Cloner.html](http://serialbasics.free.fr/Serial_Cloner.html)
  - GenBeans <http://www.genbeans.org/>
  - GeneStudio <http://genestudio.com/>
  - Seqtools <http://www.seqtools.dk/>
  - GENtle <http://gentle.magnusmanske.de/>
  - pDRAW32 <http://www.acaclone.com/>
  - DAMBE <http://dambe.bio.uottawa.ca/DAMBE/dambe.aspx>

72

## Sequence File Formats

- We have been using DNA and amino acid sequences already
- What is the typical format for these?
  - **ANSWER: Many different options**
- In order to standardize sequence data, The Nomenclature Committee of the International Union of Biochemistry and the International Union of Pure and Applied Chemistry (IUPAC) has established a standard code to represent bases that are uncertain or ambiguous.

73

## Standard Codes (IUPAC)

- IUPAC nucleotide codes and corresponding bases:
 

A = adenine	S = G or C
C = cytosine	W = A or T
G = guanine	B = G or T or C
T = thymine	D = G or A or T
U = uracil	H = A or C or T
R = G or A (purine)	V = G or C or A
Y = T or C (pyrimidine)	N = A or G or C or T (any)
K = G or T (keto)	M = A or C (amino)
M = A or C (amino)	. or - = gap
- Any other character represents an error that will not be tolerated by nearly all sequence analysis programs.

74

## Standard IUPAC Codes

- IUPAC standard single letter and three letter amino acid codes:

A Ala Alanine	F Phe Phenylalanine
R Arg Arginine	P Pro Proline
N Asn Asparagine	S Ser Serine
D Asp Aspartic acid	T Thr Threonine
C Cys Cysteine	W Trp Tryptophan
Q Gln Glutamine	Y Tyr Tyrosine
E Glu Glutamic acid	V Val Valine
G Gly Glycine	B Asx Aspartic acid or Asparagine
H His Histidine	Z Glx Glutamine or Glutamic acid
I Ile Isoleucine	X Xaa or Xxx Any amino acid
L Leu Leucine	
K Lys Lysine	
M Met Methionine	

75

## Fasta File Format

- Fasta sequence format is one of the most basic and widespread sequence formats.
- A sequence in fasta format has as its first line a descriptor beginning with a '>' character.
- The proceeding lines contain the sequence (either nucleotide or amino acid) using standard one-letter symbols.
- This format is extremely useful for sequence analysis programs, since it is devoid of numerical and non-sequence characters (with the exception of the newline character).

76

## Fasta File Format

- Example Fasta Sequence:

```
>gi|27819608|ref|NP_776342.1| hemoglobin, beta [beta globin] [Bos taurus]
MLTAEKAAVTAFWGKVKVDEVGGEALGRLLVVPYPTQRFESFGDLSTADAVMNNPKVKAHGKKVLDSF
SNGMKHLDDLKGTFAALSELHCDKLEVDPENFKLLGNVLVVVLARNFGKEFTPLQADFQKVAVGVANAL
AHRYH
```

- first line begins with '>', followed by **gi**,
  - next field surrounded by '|' is GenBank identifier
- the keyword '**ref**'
  - field will be the reference for the version of this sequence.
- final field is the description

77

## Fasta File Format

- Example Fasta Sequence:

```
>gi|27819608|ref|NP_776342.1| hemoglobin, beta [beta globin] [Bos taurus]
MLTAEKAAVTAFWGKVKVDEVGGEALGRLLVVPYPTQRFESFGDLSTADAVMNNPKVKAHGKKVLDSF
SNGMKHLDDLKGTFAALSELHCDKLEVDPENFKLLGNVLVVVLARNFGKEFTPLQADFQKVAVGVANAL
AHRYH
```

- nearly all sequence based programs treat anything following the '>' as a comment
- a few sequence analysis programs expect sequences to be in a strict fasta format

78

GenBank

- GenBank is the National Center for Biotechnology Information’s nucleic acid and protein sequence database.
- It is the most widely used source of biological sequence data.
- GenBank file format contains information about the sequence, including literature references, functions of the sequence, locations of various features, etc.
- information organized into fields, each with an identifier, justified to the farthest left column.
- Some identifiers have additional subfields.
- sequence data lies between the identifier ORIGIN and the “//” which signals the end of a GenBank record.

79

GenBank Record

LOCUS HBB 145 aa linear MAM 22-JAN-2003  
DEFINITION hemoglobin, beta [beta globin] [Bos taurus].  
ACCESSION NP\_776342  
VERSION NP\_776342.1 GI:27819608  
DBSOURCE REFSEQ: accession [NM\\_173917.1](#)  
KEYWORDS .  
SOURCE Bos taurus (cow)  
ORGANISM [Bos taurus](#); Eukaryota; Metazoa; Chordata; Craniata;  
Vertebrata; Euteleostomi; Mammalia; Eutheria; Cetartiodactyla;  
Ruminantia; Pecora; Bovidae; Bovinae; Bos.  
REFERENCE 1 (residues 1 to 145)  
AUTHORS Duncan,C.H.  
JOURNAL Unpublished (1991)  
COMMENT PROVISIONAL [REFSEQ](#): This record has not yet been subject to final  
NCBI review. The reference sequence was derived from [M63453.1](#).  
FEATURES Location/Qualifiers source 1..145

80

ASN.1

- Abstract Syntax Notation (ASN.1):
  - formal description language developed to encode various data to be easily connected across computer systems
- ASN.1 is highly structured and detailed
- ASN.1 format contains all of the other information found in other formats

81