

The Bioinformatics Approach to Proteins

Magnus Andersson

magnus.andersson@scilifelab.se

Theoretical & Computational Biophysics

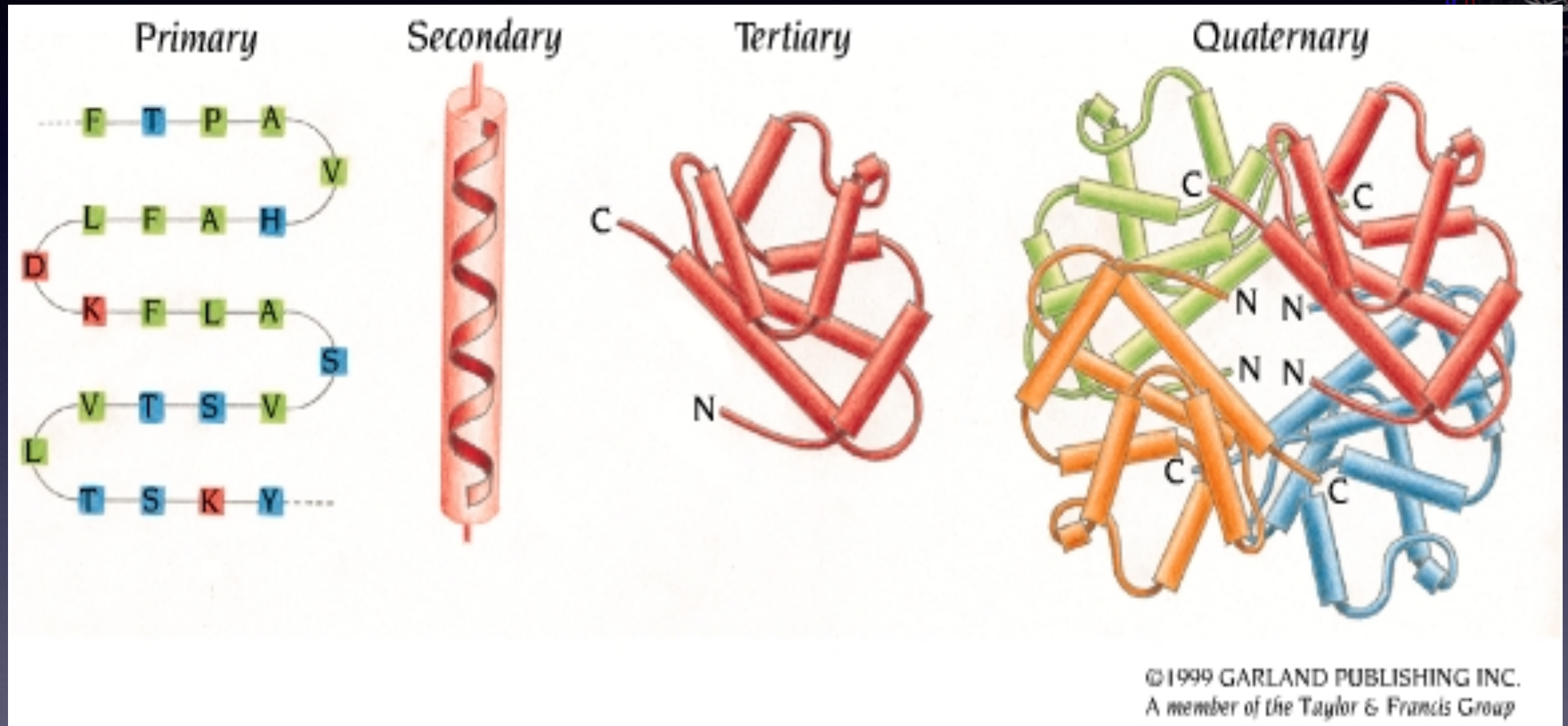
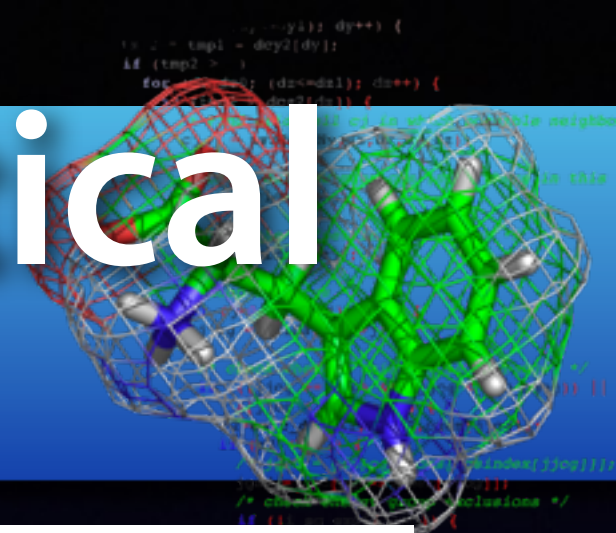
SciLifeLab



[illegible]

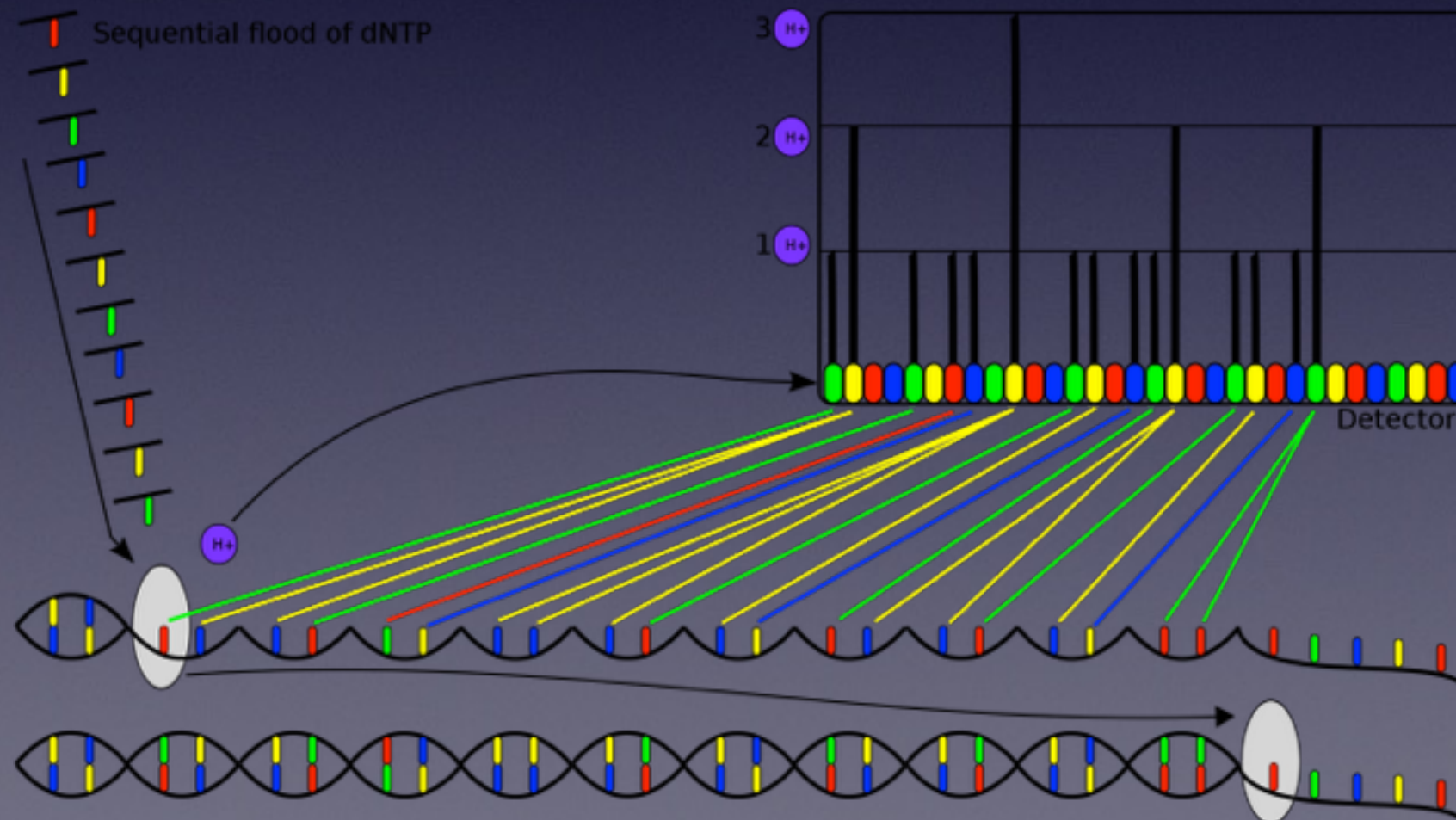
- Genomes, genes & evolution
- Large scale databases
- Sequence comparison, finding genes
- Sequence - structure - function
- Evolution vs. laws of nature
 - *Computer science vs. chemistry/physics?*

Intellectual & practical problems



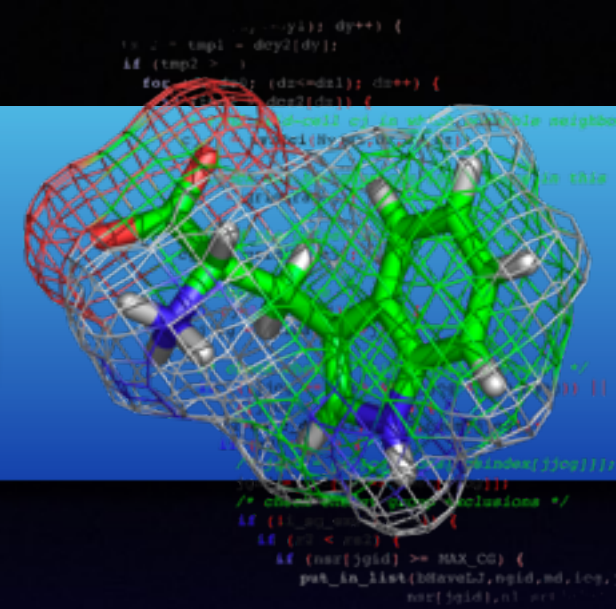
*It is interesting to understand **how** structure forms,
but it would also be worth a lot if we could
just **predict** the final structure!*

DNA sequencing



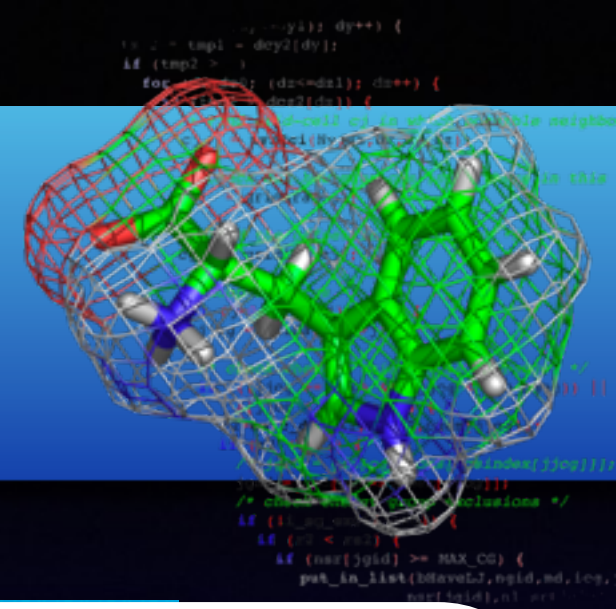
```
/* ... */  
if (i1 == 0) {  
    if (i2 < 0) {  
        if (i3 > 0) {  
            get_in_list(i1, i2, i3, i4, i5, i6, i7, i8, i9, i10, i11, i12, i13, i14, i15, i16, i17, i18, i19, i20, i21, i22, i23, i24, i25, i26, i27, i28, i29, i30, i31, i32, i33, i34, i35, i36, i37, i38, i39, i40, i41, i42, i43, i44, i45, i46, i47, i48, i49, i50, i51, i52, i53, i54, i55, i56, i57, i58, i59, i60, i61, i62, i63, i64, i65, i66, i67, i68, i69, i70, i71, i72, i73, i74, i75, i76, i77, i78, i79, i80, i81, i82, i83, i84, i85, i86, i87, i88, i89, i90, i91, i92, i93, i94, i95, i96, i97, i98, i99, i100, i101, i102, i103, i104, i105, i106, i107, i108, i109, i110, i111, i112, i113, i114, i115, i116, i117, i118, i119, i120, i121, i122, i123, i124, i125, i126, i127, i128, i129, i130, i131, i132, i133, i134, i135, i136, i137, i138, i139, i140, i141, i142, i143, i144, i145, i146, i147, i148, i149, i150, i151, i152, i153, i154, i155, i156, i157, i158, i159, i160, i161, i162, i163, i164, i165, i166, i167, i168, i169, i170, i171, i172, i173, i174, i175, i176, i177, i178, i179, i180, i181, i182, i183, i184, i185, i186, i187, i188, i189, i190, i191, i192, i193, i194, i195, i196, i197, i198, i199, i200, i201, i202, i203, i204, i205, i206, i207, i208, i209, i210, i211, i212, i213, i214, i215, i216, i217, i218, i219, i220, i221, i222, i223, i224, i225, i226, i227, i228, i229, i230, i231, i232, i233, i234, i235, i236, i237, i238, i239, i240, i241, i242, i243, i244, i245, i246, i247, i248, i249, i250, i251, i252, i253, i254, i255, i256, i257, i258, i259, i260, i261, i262, i263, i264, i265, i266, i267, i268, i269, i270, i271, i272, i273, i274, i275, i276, i277, i278, i279, i280, i281, i282, i283, i284, i285, i286, i287, i288, i289, i290, i291, i292, i293, i294, i295, i296, i297, i298, i299, i300, i301, i302, i303, i304, i305, i306, i307, i308, i309, i310, i311, i312, i313, i314, i315, i316, i317, i318, i319, i320, i321, i322, i323, i324, i325, i326, i327, i328, i329, i330, i331, i332, i333, i334, i335, i336, i337, i338, i339, i340, i341, i342, i343, i344, i345, i346, i347, i348, i349, i350, i351, i352, i353, i354, i355, i356, i357, i358, i359, i360, i361, i362, i363, i364, i365, i366, i367, i368, i369, i370, i371, i372, i373, i374, i375, i376, i377, i378, i379, i380, i381, i382, i383, i384, i385, i386, i387, i388, i389, i390, i391, i392, i393, i394, i395, i396, i397, i398, i399, i400, i401, i402, i403, i404, i405, i406, i407, i408, i409, i410, i411, i412, i413, i414, i415, i416, i417, i418, i419, i420, i421, i422, i423, i424, i425, i426, i427, i428, i429, i430, i431, i432, i433, i434, i435, i436, i437, i438, i439, i440, i441, i442, i443, i444, i445, i446, i447, i448, i449, i450, i451, i452, i453, i454, i455, i456, i457, i458, i459, i460, i461, i462, i463, i464, i465, i466, i467, i468, i469, i470, i471, i472, i473, i474, i475, i476, i477, i478, i479, i480, i481, i482, i483, i484, i485, i486, i487, i488, i489, i490, i491, i492, i493, i494, i495, i496, i497, i498, i499, i500, i501, i502, i503, i504, i505, i506, i507, i508, i509, i510, i511, i512, i513, i514, i515, i516, i517, i518, i519, i520, i521, i522, i523, i524, i525, i526, i527, i528, i529, i530, i531, i532, i533, i534, i535, i536, i537, i538, i539, i540, i541, i542, i543, i544, i545, i546, i547, i548, i549, i550, i551, i552, i553, i554, i555, i556, i557, i558, i559, i560, i561, i562, i563, i564, i565, i566, i567, i568, i569, i570, i571, i572, i573, i574, i575, i576, i577, i578, i579, i580, i581, i582, i583, i584, i585, i586, i587, i588, i589, i590, i591, i592, i593, i594, i595, i596, i597, i598, i599, i600, i601, i602, i603, i604, i605, i606, i607, i608, i609, i610, i611, i612, i613, i614, i615, i616, i617, i618, i619, i620, i621, i622, i623, i624, i625, i626, i627, i628, i629, i630, i631, i632, i633, i634, i635, i636, i637, i638, i639, i640, i641, i642, i643, i644, i645, i646, i647, i648, i649, i650, i651, i652, i653, i654, i655, i656, i657, i658, i659, i660, i661, i662, i663, i664, i665, i666, i667, i668, i669, i670, i671, i672, i673, i674, i675, i676, i677, i678, i679, i680, i681, i682, i683, i684, i685, i686, i687, i688, i689, i690, i691, i692, i693, i694, i695, i696, i697, i698, i699, i700, i701, i702, i703, i704, i705, i706, i707, i708, i709, i710, i711, i712, i713, i714, i715, i716, i717, i718, i719, i720, i721, i722, i723, i724, i725, i726, i727, i728, i729, i730, i731, i732, i733, i734, i735, i736, i737, i738, i739, i740, i741, i742, i743, i744, i745, i746, i747, i748, i749, i750, i751, i752, i753, i754, i755, i756, i757, i758, i759, i760, i761, i762, i763, i764, i765, i766, i767, i768, i769, i770, i771, i772, i773, i774, i775, i776, i777, i778, i779, i780, i781, i782, i783, i784, i785, i786, i787, i788, i789, i790, i791, i792, i793, i794, i795, i796, i797, i798, i799, i800, i801, i802, i803, i804, i805, i806, i807, i808, i809, i810, i811, i812, i813, i814, i815, i816, i817, i818, i819, i820, i821, i822, i823, i824, i825, i826, i827, i828, i829, i830, i831, i832, i833, i834, i835, i836, i837, i838, i839, i840, i841, i842, i843, i844, i845, i846, i847, i848, i849, i850, i851, i852, i853, i854, i855, i856, i857, i858, i859, i860, i861, i862, i863, i864, i865, i866, i867, i868, i869, i870, i871, i872, i873, i874, i875, i876, i877, i878, i879, i880, i881, i882, i883, i884, i885, i886, i887, i888, i889, i890, i891, i892, i893, i894, i895, i896, i897, i898, i899, i900, i901, i902, i903, i904, i905, i906, i907, i908, i909, i910, i911, i912, i913, i914, i915, i916, i917, i918, i919, i920, i921, i922, i923, i924, i925, i926, i927, i928, i929, i930, i931, i932, i933, i934, i935, i936, i937, i938, i939, i940, i941, i942, i943, i944, i945, i946, i947, i948, i949, i950, i951, i952, i953, i954, i955, i956, i957, i958, i959, i960, i961, i962, i963, i964, i965, i966, i967, i968, i969, i970, i971, i972, i973, i974, i975, i976, i977, i978, i979, i980, i981, i982, i983, i984, i985, i986, i987, i988, i989, i990, i991, i992, i993, i994, i995, i996, i997, i998, i999, i1000, i1001, i1002, i1003, i1004, i1005, i1006, i1007, i1008, i1009, i1010, i1011, i1012, i1013, i1014, i1015, i1016, i1017, i1018, i1019, i1020, i1021, i1022, i1023, i1024, i1025, i1026, i1027, i1028, i1029, i1030, i1031, i1032, i1033, i1034, i1035, i1036, i1037, i1038, i1039, i1040, i1041, i1042, i1043, i1044, i1045, i1046, i1047, i1048, i1049, i1050, i1051, i1052, i1053, i1054, i1055, i1056, i1057, i1058, i1059, i1060, i1061, i1062, i1063, i1064, i1065, i1066, i1067, i1068, i1069, i1070, i1071, i1072, i1073, i1074, i1075, i1076, i1077, i1078, i1079, i1080, i1081, i1082, i1083, i1084, i1085, i1086, i1087, i1088, i1089, i1090, i1091, i1092, i1093, i1094, i1095, i1096, i1097, i1098, i1099, i1100, i1101, i1102, i1103, i1104, i1105, i1106, i1107, i1108, i1109, i1110, i1111, i1112, i1113, i1114, i1115, i1116, i1117, i1118, i1119, i1120, i1121, i1122, i1123, i1124, i1125, i1126, i1127, i1128, i1129, i1130, i1131, i1132, i1133, i1134, i1135, i1136, i1137, i1138, i1139, i1140, i1141, i1142, i1143, i1144, i1145, i1146, i1147, i1148, i1149, i1150, i1151, i1152, i1153, i1154, i1155, i1156, i1157, i1158, i1159, i1160, i1161, i1162, i1163, i1164, i1165, i1166, i1167, i1168, i1169, i1170, i1171, i1172, i1173, i1174, i1175, i1176, i1177, i1178, i1179, i1180, i1181, i1182, i1183, i1184, i1185, i1186, i1187, i1188, i1189, i1190, i1191, i1192, i1193, i1194, i1195, i1196, i1197, i1198, i1199, i1200, i1201, i1202, i1203, i1204, i1205, i1206, i1207, i1208, i1209, i1210, i1211, i1212, i1213, i1214, i1215, i1216, i1217, i1218, i1219, i1220, i1221, i1222, i1223, i1224, i1225, i1226, i1227, i1228, i1229, i1230, i1231, i1232, i1233, i1234, i1235, i1236, i1237, i1238, i1239, i1240, i1241, i1242, i1243, i1244, i1245, i1246, i1247, i1248, i1249, i1250, i1251, i1252, i1253, i1254, i1255, i1256, i1257, i1258, i1259, i1260, i1261, i1262, i1263, i1264, i1265, i1266, i1267, i1268, i1269, i1270, i1271, i1272, i1273, i1274, i1275, i1276, i1277, i1278, i1279, i1280, i1281, i1282, i1283, i1284, i1285, i1286, i1287, i1288, i1289, i1290, i1291, i1292, i1293, i1294, i1295, i1296, i1297, i1298, i1299, i1300, i1301, i1302, i1303, i1304, i1305, i1306, i1307, i1308, i1309, i1310, i1311, i1312, i1313, i1314, i1315, i1316, i1317, i1318, i1319, i1320, i1321, i1322, i1323, i1324, i1325, i1326, i1327, i1328, i1329, i1330, i1331, i1332, i1333, i1334, i1335, i1336, i1337, i1338, i1339, i1340, i1341, i1342, i1343, i1344, i1345, i1346, i1347, i1348, i1349, i1350, i1351, i1352, i1353, i1354, i1355, i1356, i1357, i1358, i1359, i1360, i1361, i1362, i1363, i1364, i1365, i1366, i1367, i1368, i1369, i1370, i1371, i1372, i1373, i1374, i1375, i1376, i1377, i1378, i1379, i1380, i1381, i1382, i1383, i1384, i1385, i1386, i1387, i1388, i1389, i1390, i1391, i1392, i1393, i1394, i1395, i1396, i1397, i1398, i1399, i1400, i1401, i1402, i1403, i1404, i1405, i1406, i1407, i1408, i1409, i1410, i1411, i1412, i1413, i1414, i1415, i1416, i1417, i1418, i1419, i1420, i1421, i1422, i1423, i1424, i1425, i1426, i1427, i1428, i1429, i1430, i1431, i1432, i1433, i1434, i1435, i1436, i1437, i1438, i1439, i1440, i1441, i1442, i1443, i1444, i1445, i1446, i1447, i1448, i1449, i1450, i1451, i1452, i1453, i1454, i1455, i1456, i1457, i1458, i1459, i1460, i1461, i1462, i1463, i1464, i1465, i1466, i1467, i1468, i1469, i1470, i1471, i1472, i1473, i1474, i1475, i1476, i1477, i1478, i1479, i1480, i1481, i1482, i1483, i1484, i1485, i1486, i1487, i1488, i1489, i1490, i1491, i1492, i1493, i1494, i1495, i1496, i1497, i1498, i1499, i1500, i1501, i1502, i1503, i1504, i1505, i1506, i1507, i1508, i1509, i1510, i1511, i1512, i1513, i1514, i1515, i1516, i1517, i1518, i1519, i1520, i1521, i1522, i1523, i1524, i1525, i1526, i1527, i1528, i1529, i1530, i1531, i1532, i1533, i1534, i1535, i1536, i1537, i1538, i1539, i1540, i1541, i1542, i1543, i1544, i1545, i1546, i1547, i1548, i1549, i1550, i1551, i1552, i1553, i1554, i1555, i1556, i1557, i1558, i1559, i1560, i1561, i1562, i1563, i1564, i1565, i1566, i1567, i1568, i1569, i1570, i1571, i1572, i1573, i1574, i1575, i1576, i1577, i1578, i1579, i1580, i1581, i1582, i1583, i1584, i1585, i1586, i1587, i1588, i1589, i1590, i1591, i1592, i1593, i1594, i1595, i1596, i1597, i1598, i1599, i1600, i1601, i1602, i1603, i1604, i1605, i1606, i1607, i1608, i1609, i1610, i1611, i1612, i1613, i1614, i1615, i1616, i1617, i1618, i1619, i1620, i1621, i1622, i1623, i1624, i1625, i1626, i1627, i1628, i1629, i1630, i1631, i1632, i1633, i1634, i1635, i1636, i1637, i1638, i1639, i1640, i1641, i1642, i1643, i1644, i1645, i1646, i1647, i1648, i1649, i1650, i1651, i1652, i1653, i1654, i1655, i1656, i1657, i1658, i1659, i1660, i1661, i1662, i1663, i1664, i1665, i1666, i1667, i1668, i1669, i1670, i1671, i1672, i1673, i1674, i1675, i1676, i1677, i1678, i1679, i1680, i1681, i1682, i1683, i1684, i1685, i1686, i1687, i1688, i1689, i1690, i1691, i1692, i1693, i1694, i1695, i1696, i1697, i1698, i1699, i1700, i1701, i1702, i1703, i1704, i1705, i1706, i1707, i1708, i1709, i1710, i1711, i1712, i1713, i1714, i1715, i1716, i1717, i1718, i1719, i1720, i1721, i1722, i1723, i1724, i1725, i1726, i1727, i1728, i1729, i1730, i1731, i1732, i1733, i1734, i1735, i1736, i1737, i1738, i1739, i1740, i1741, i1742, i1743, i1744, i1745, i1746, i1747, i1748, i1749, i1750, i1751, i1752, i1753, i1754, i1755, i1756, i1757, i1758, i1759, i1760, i1761, i1762, i1763, i1764, i1765, i1766, i1767, i1768, i1769, i1770, i1771, i1772, i1773, i1774, i1775, i1776, i1777, i1778, i1779, i1780, i1781, i1782, i1783, i1784, i1785, i1786, i1787, i1788, i1789, i1790, i1791, i1792, i1793, i1794, i1795, i1796, i1797, i1798, i1799, i1800, i1801, i1802, i1803, i1804, i1805, i1806, i1807, i1808, i1809, i1810, i1811, i1812, i1813, i1814, i1815, i1816, i1817, i1818, i1819, i1820, i1821, i1822, i1823, i1824, i1825, i1826, i1827, i1828, i1829, i1830, i1831, i1832, i1833, i1834, i1835, i1836, i1837, i1838, i1839, i1840, i1841, i1842, i1843, i1844, i1845, i1846, i1847, i1848, i1849, i1850, i1851, i1852, i1853, i1854, i1855, i1856, i1857, i1858, i1859, i1860, i1861, i1862, i1863, i1864, i1865, i1866, i1867, i1868, i1869, i1870, i1871, i1872, i1873, i1874, i1875, i1876, i1877, i1878, i1879, i1880, i1881, i1882, i1883, i1884, i1885, i1886, i1887, i1888, i1889, i1890, i1891, i1892, i1893, i1894, i1895, i1896, i1897, i1898, i1899, i1900, i1901, i1902, i1903, i1904, i1905, i1906, i1907, i1908, i1909, i1910, i1911, i1912, i1913, i1914, i1915, i1916, i1917, i1918, i1919, i1920, i1921, i1922, i1923, i1924, i1925, i1926, i1927, i1928, i1929, i1930, i1931, i1932, i1933, i1934, i1935, i1936, i1937, i1938, i1939, i1940, i1941, i1942, i1943, i1944, i1945, i1946, i1947, i1948, i1949, i1950, i1951, i1952, i1953, i1954, i1955, i1956, i1957, i1958, i1959, i1960, i1961, i1962, i1963, i1964, i1965, i1966, i1967, i1968, i1969, i1970, i1971, i1972, i1973, i1974, i1975, i1976, i1977, i1978, i1979, i1980, i1981, i1982, i1983, i1984, i1985, i1986, i1987, i1988, i1989, i1990, i1991, i1992, i1993, i1994, i1995, i1996, i1997, i1998, i1999, i2000, i2001, i2002, i2003, i2004, i2005, i2006, i2007, i2008, i2009, i2010, i2011, i2012, i2013, i2014, i2015, i2016, i2017, i2018, i2019, i2020, i2021, i2022, i2023, i2024, i2025, i2026, i2027, i2028, i2029, i2030, i2031, i2032, i2033, i2034, i2035, i2036, i2037, i2038, i2039, i2040, i2041, i2042, i2043, i2044, i2045, i2046, i2047, i2048, i2049, i2050, i2051, i2052, i2053, i2054, i2055, i2056, i2057, i2058, i2059, i2060, i2061, i2062, i2063, i2064, i2065, i2066, i2067, i2068, i2069, i2070, i2071, i2072, i2073, i2074, i2075, i2076, i2077, i2078, i2079, i2080, i2081, i2082, i2083, i2084, i2085, i2086, i2087, i2088, i2089, i2090, i2091, i2092, i2093, i2094, i2095, i2096, i2097, i2098, i2099, i2100, i2101, i2102, i2103, i2104, i2105, i2106, i2107, i2108, i2109, i2110, i2111, i2112, i2113, i2114, i2115, i2116, i2117, i2118, i2119, i2120, i2121, i2122, i2123, i2124, i2125, i2126, i2127, i2128, i2129, i2130, i2131, i2132, i2133, i2134, i2135, i2136, i2137, i2138, i2139, i2140, i2141, i2142, i2143, i2144, i2145, i2146, i2147, i2148, i2149, i2150, i2151, i2152, i2153, i2154, i2155, i2156, i2157, i2158, i2159, i2160, i2161, i2162, i2163, i2164, i2165, i2166, i216
```


DNA vs protein



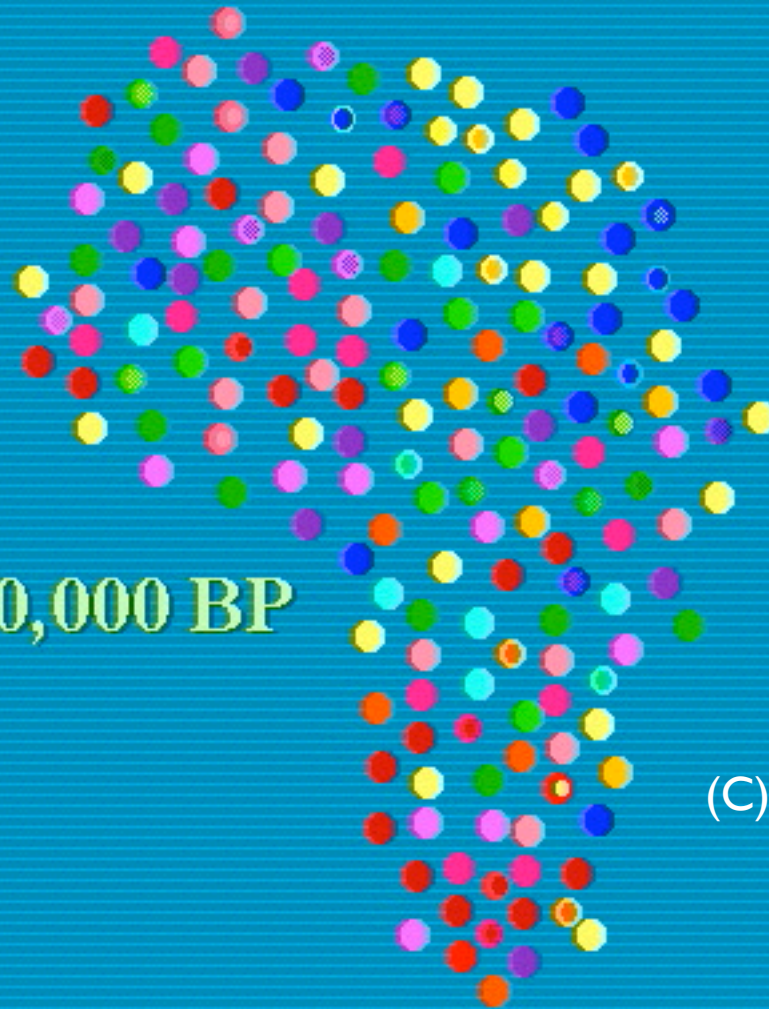
- 1.2% protein-coding DNA in human
 - ORF: Open Reading Frame
 - ATG TAA
- 20,000-25,000 genes in human
- How do we find & study similarities?

Human evolution



Early Homo sapiens sapiens
in Africa

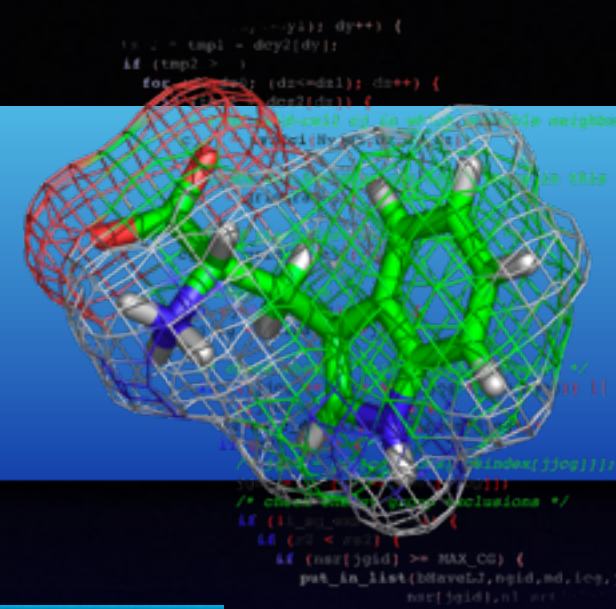
150,000 to 100,000 BP



BP=Before Present

(C) Kenneth Kidd, Yale University

Human evolution



Homo sapiens sapiens
colonizing south west Asia

(C) Kenneth Kidd, Yale University
~100,000 BP

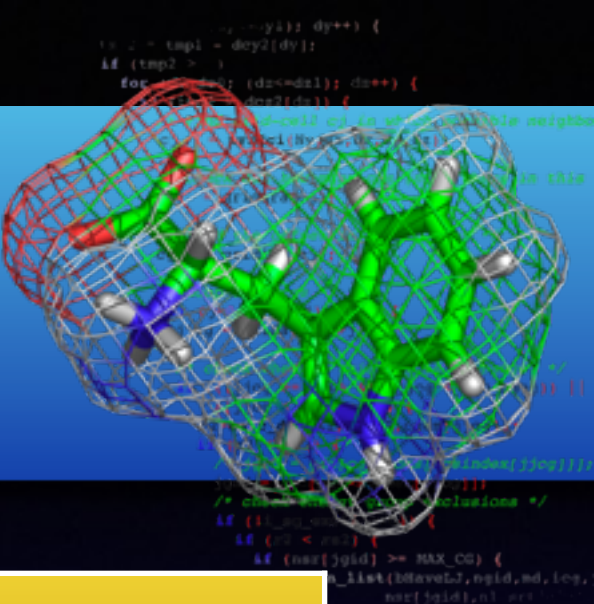
[illegible]

(C) Kenneth Kidd, Yale University

[illegible]

- **BRCA1/BRCA2 (=BReast CAncer)**
- **Some DNA mutations in these mean 85% risk of developing breast cancer**
- **New efficient genetic tests for screening**
 - **Frequent mamograms if positive**
 - **Possibly preventive breast removal**

Nucleotides determine the amino acid sequence



	T	C	A	G	
T	Phe Phe Leu Leu	Ser Ser Ser Ser	Tyr Tyr STOP STOP	Cys Cys STOP Trp	T C A G
2C	Leu Leu Leu Leu	Pro Pro Pro Pro	His His Gln Gln	Arg Arg Arg Arg	T C A G3
A	Ile Ile Ile Met	Thr Thr Thr Thr	Asn Asn Lys Lys	Ser Ser Arg Arg	T C A G
G	Val Val Val Val	Ala Ala Ala Ala	Asp Asp Glu Glu	Gly Gly Gly Gly	T C A G

1	KIEEGKLVIW	INGDKGYNGL	AEVGKKFEKD	TGIKVTVEHP
41	DKLEEKFPQV	AATGDGPDII	FWAHDRFGGY	AQSGLLAEIT
81	PDKAFQDKLY	PFTWDAVRYN	GKLIAYPIAV	EALSLIYNKD
121	LLPNPPKTWE	EIPALDKELK	AKGKSALMFN	LQEPYFTWPL
161	IAADGGYAFK	YENGKYDIKD	VGVDNAGAKA	GLTFLVDLIK
201	NKHMNADTDY	SIAEAAFNKG	ETAMTINGPW	AWSNIDTSKV
241	NYGVTVLPTF	KGQPSKPFVG	VLSAGINAAS	PNKELAKEFL
301	ENYLLTDEGL	EAVNKDKPLG	AVALKSYEEE	LAKDPRIAAT
341	MENAQKGEIM	PNIPQMSAFW	YAVRTAVINA	ASGRQTVDEA
361	LKDAQTRITK			



A 3D ribbon diagram of a protein-ligand complex. The protein is composed of several subunits, each represented by a different color: blue, yellow, orange, green, and red. The subunits are interconnected by various loops and helices. In the center of the protein structure, there is a cluster of red spheres, representing a ligand or a small molecule. The background is black.

Ligand Binding

**Feedback to sequence:
Natural Selection**

Sequence

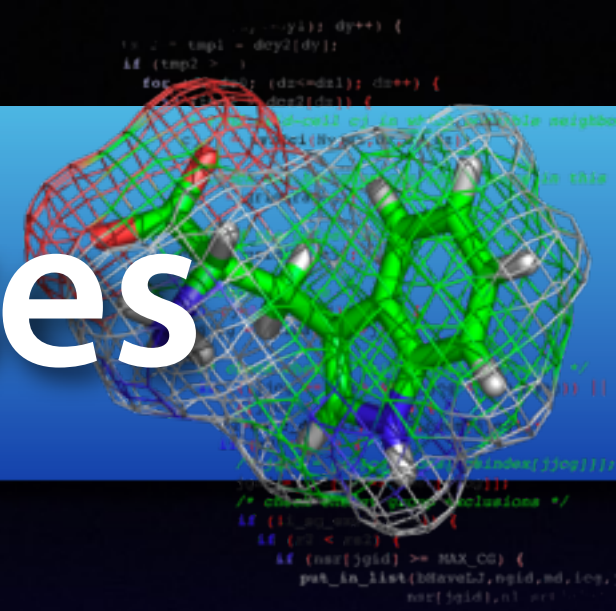
Structure

Function

[illegible]

-
- Bases**
- 1,000,000,000,000
100,000,000,000
10,000,000,000
1,000,000,000
100,000,000
10,000,000
1,000,000
- 1985 1990 1995 2000 2005 2010
- GenBank
WGS

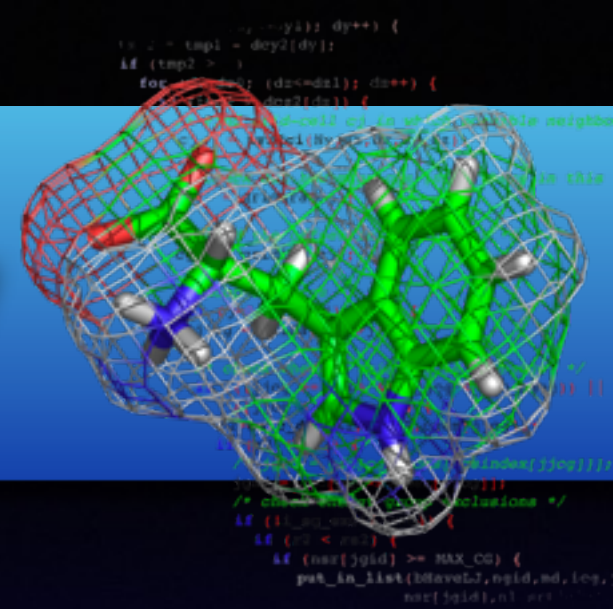
Some Public Databases



- GenBank (NCBI) - genome sequences
 - Huge, but lots of junk
- SwissProt/TrEMBL - Annotated seqs.
 - Genes known to code for proteins
- Protein Data Bank (PDB)
 - Coordinates of 3D protein structures

[illegible]

Sequence Similarity



- Natural selection:
 - Random mutation/insertion/deletion
 - Survival of the fittest
- Evolution from older ancestors
- Proteins (genes) from a common ancestor are called *Homologs*

[illegible]

- ***Paralogs***: Homologous proteins that perform different (but related) functions in the same organism
- ***Orthologs***: Homologous proteins that perform the same (or very similar) function in different organisms

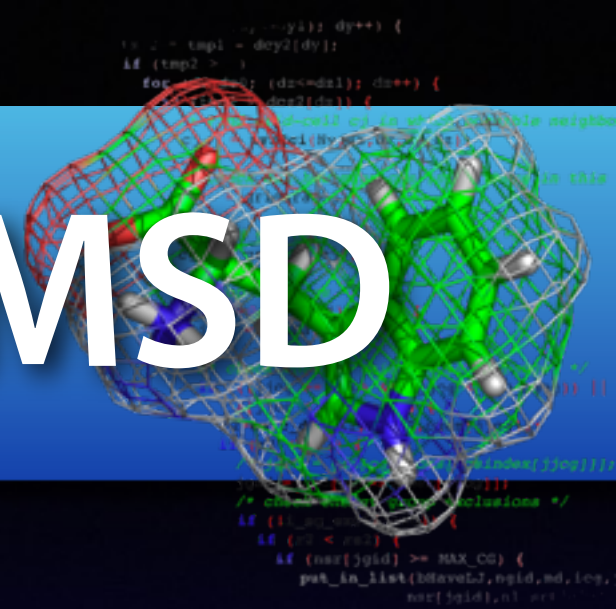
Myoglobin from 9 species



Are these paralogs or orthologs?

MYHU	. . M	GLSDGEWQLVLNV	WGKVEAD	IPGH	GQEV	LIRL	FKGH	PE . .
MYCZ	. . .	GLSDGEWQLVLNV	WGKVEAD	IPGH	GQEV	LIRL	FKGH	PE . .
MYMQV	. . .	GLSDGEWQLVLNI	WGKVEAD	IPSH	GQEV	LISL	FKGH	PE . .
MYOY	. . .	GLSDAEWQLVLNV	WGKVEAD	IPGH	GQD	VLIR	FKGH	PE . .
MYFXBE	. . .	GLSDGEWQIVLNI	WGKVETD	LAGH	GQEV	LIRL	FKN	HPE . .
MYDG	. . .	GLSDGEWQIVLNI	WGKVETD	LAGH	GQEV	LIRL	FKN	HPE . .
MYWHL	. . .	GLSDGEWQLVLNV	WGKVEAD	LAGH	GQD	ILIR	FKGH	PE . .
MYPN	. . .	GLNDQEWQQV	LTMWGKVESD	LAGH	GHAV	LMRL	FKS	HPE . .
MYTUY	ADFDA	VLKCV	GPVEAD	YT	TMGGL	VLTR	LFKEHPE . .
Consensus		GLSDG	ewQL	N	K	A	GH	QEv IR G

Structure distance: RMSD

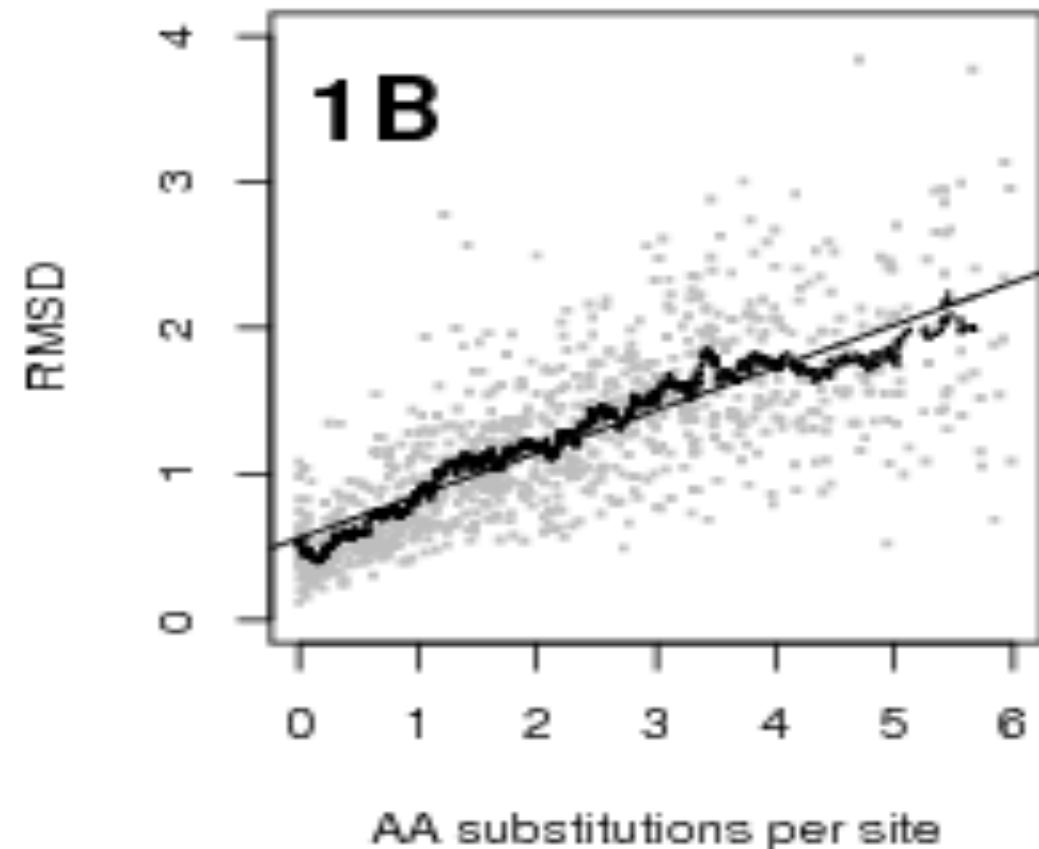
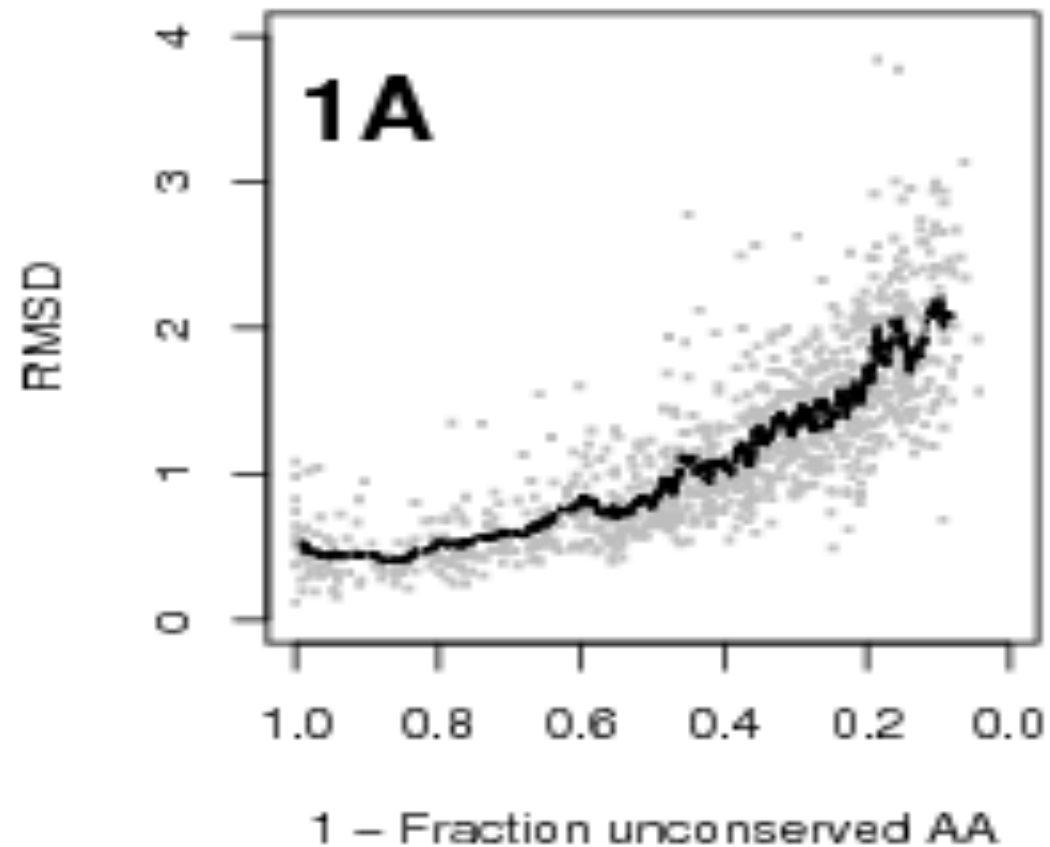


- Defined almost like a standard deviation

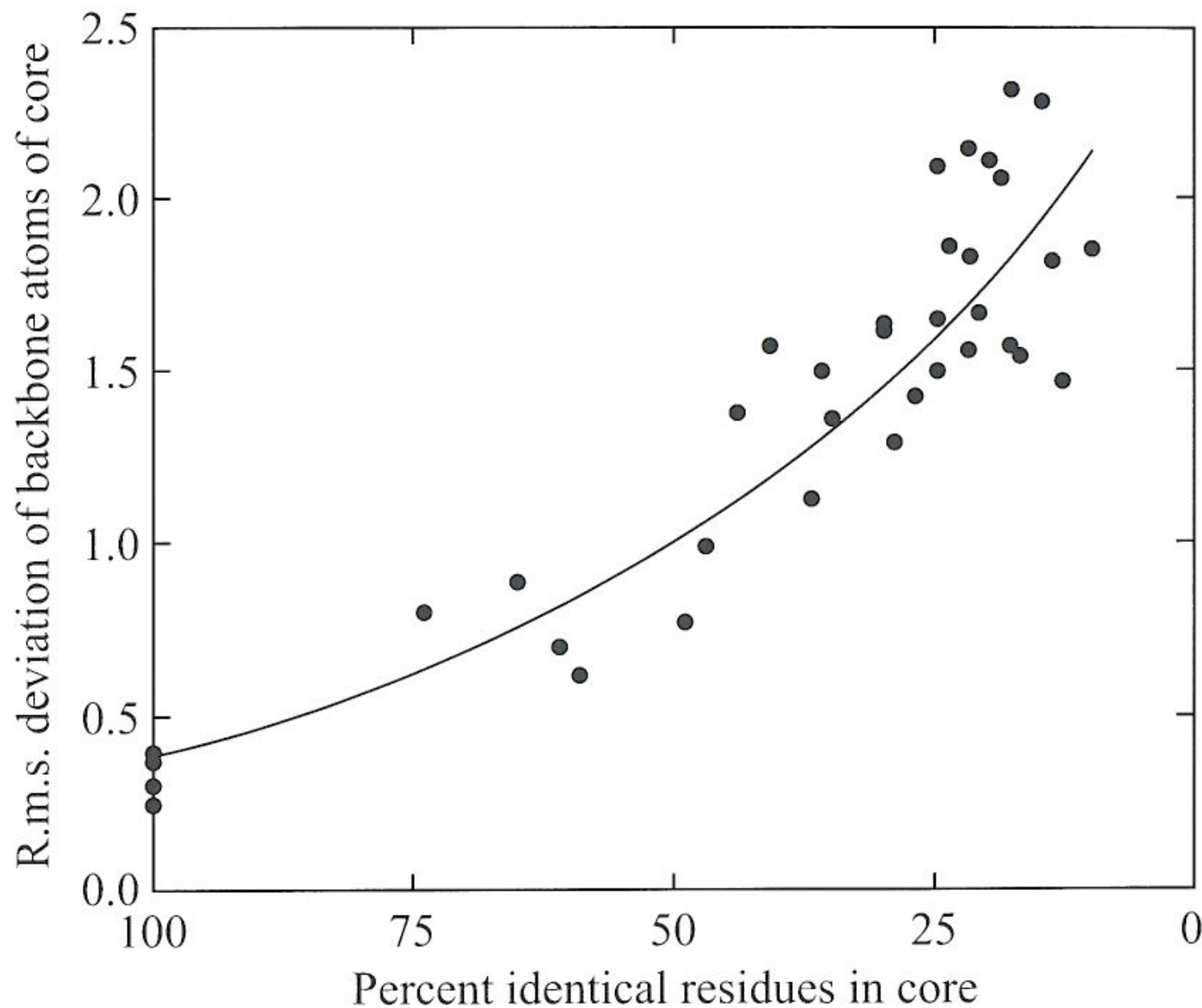
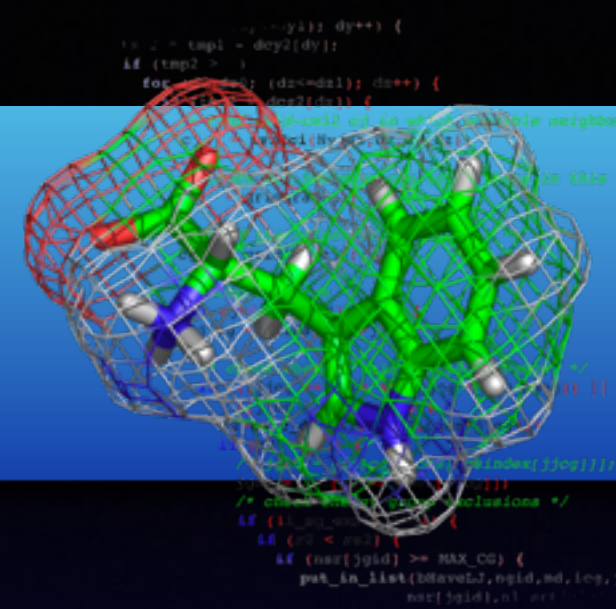
$$\sum_{i=1}^n \sqrt{\frac{(x_a - x_b)^2 + (y_a - y_b)^2 + (z_a - z_b)^2}{n}}$$

- Average displacement of atoms
- X-ray: 0.2 Å NMR: 1-2 Å
- Homology models: 1-3 Å

Structural change depends on evolutionary distance!



Homology is useful for structure prediction

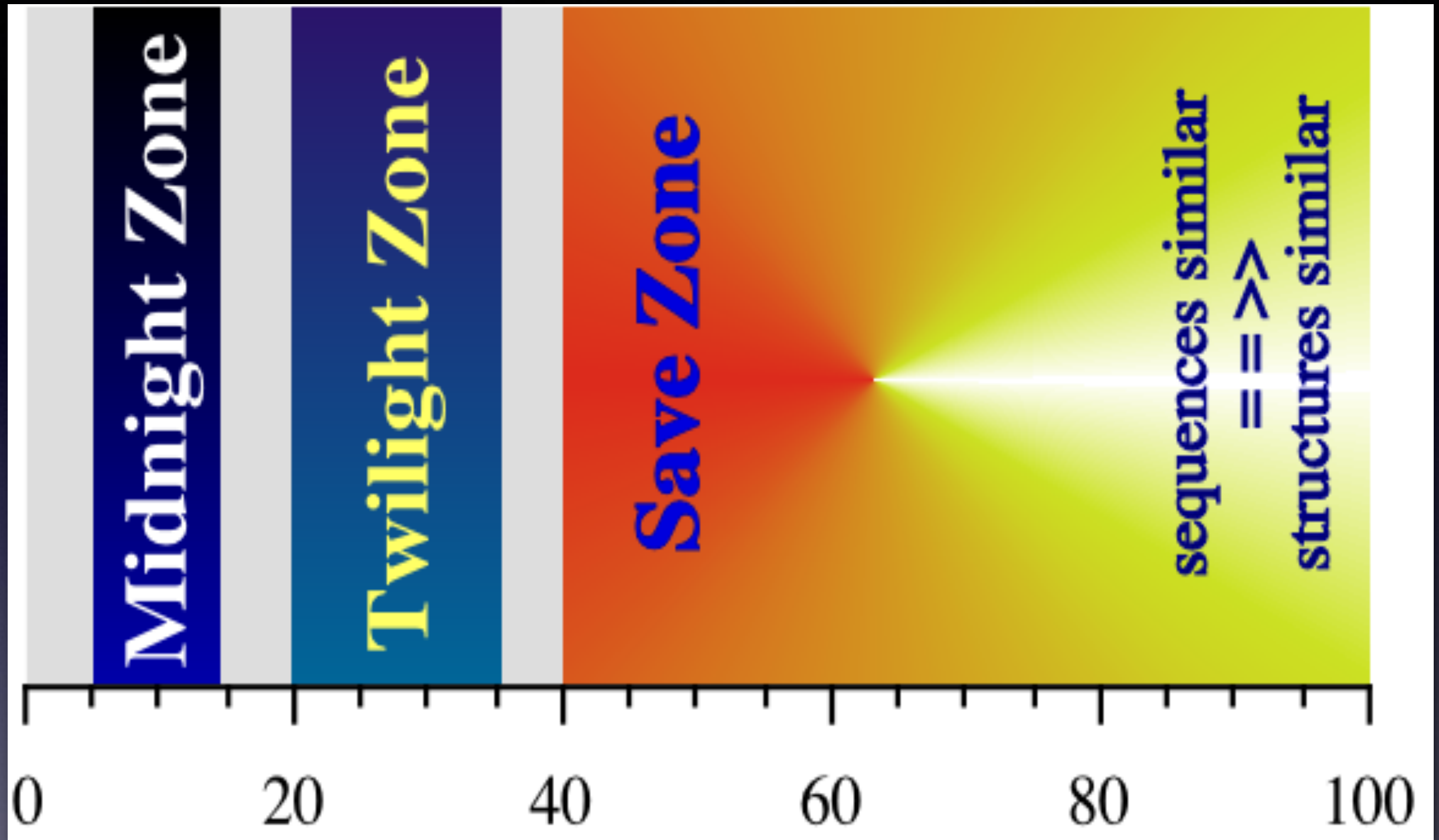


If we know the structure of a homologous protein, we might be able to build a model based on this relative!



Impossible *Hard*

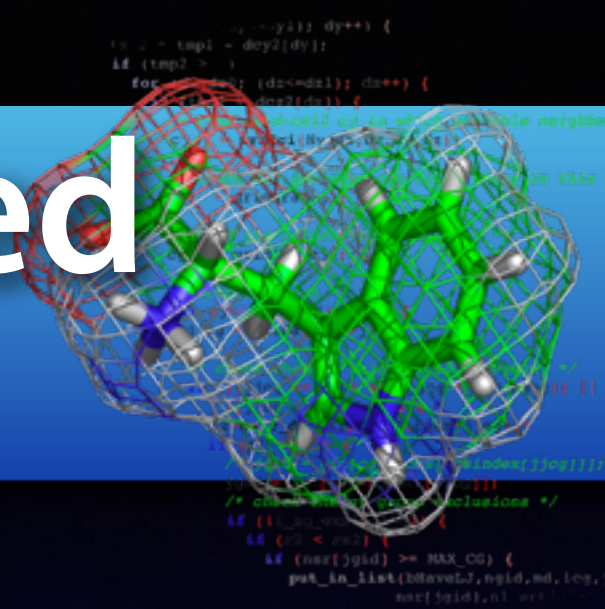
Easy



Sequence identity

But: Proteins are either homologs or not - the question is only when we can detect it! (You can't be 50% siblings)

Homology can be detected from sequence similarity



- How do we locate & assess similarities?
- Alignment of sequences (just line up?)

ACKFLFGDELR
CKFARLFADEL



ACKF--LFGDELR
CKFARLFADEL

- What do we do with mismatches?
- Insertions? Deletions? Ends?

Mismatch

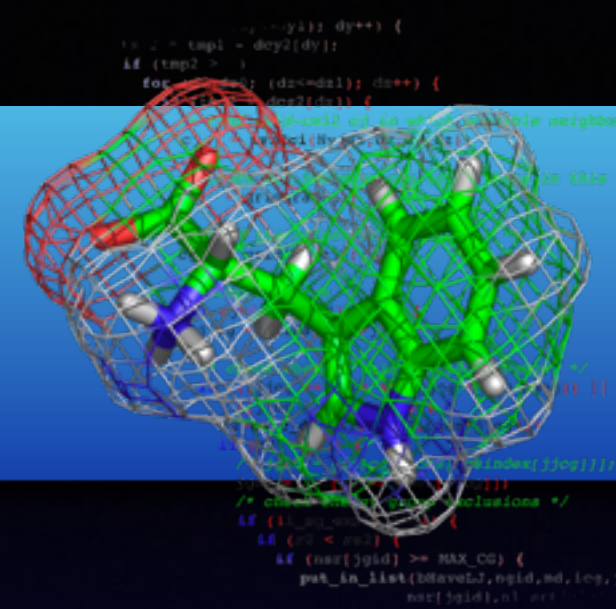
Insertion

	A	C	K	F	L	F	G	D	E	L	R
C											
K											
F											
A											
R											
L											
F											
G											
D											
E											
L											

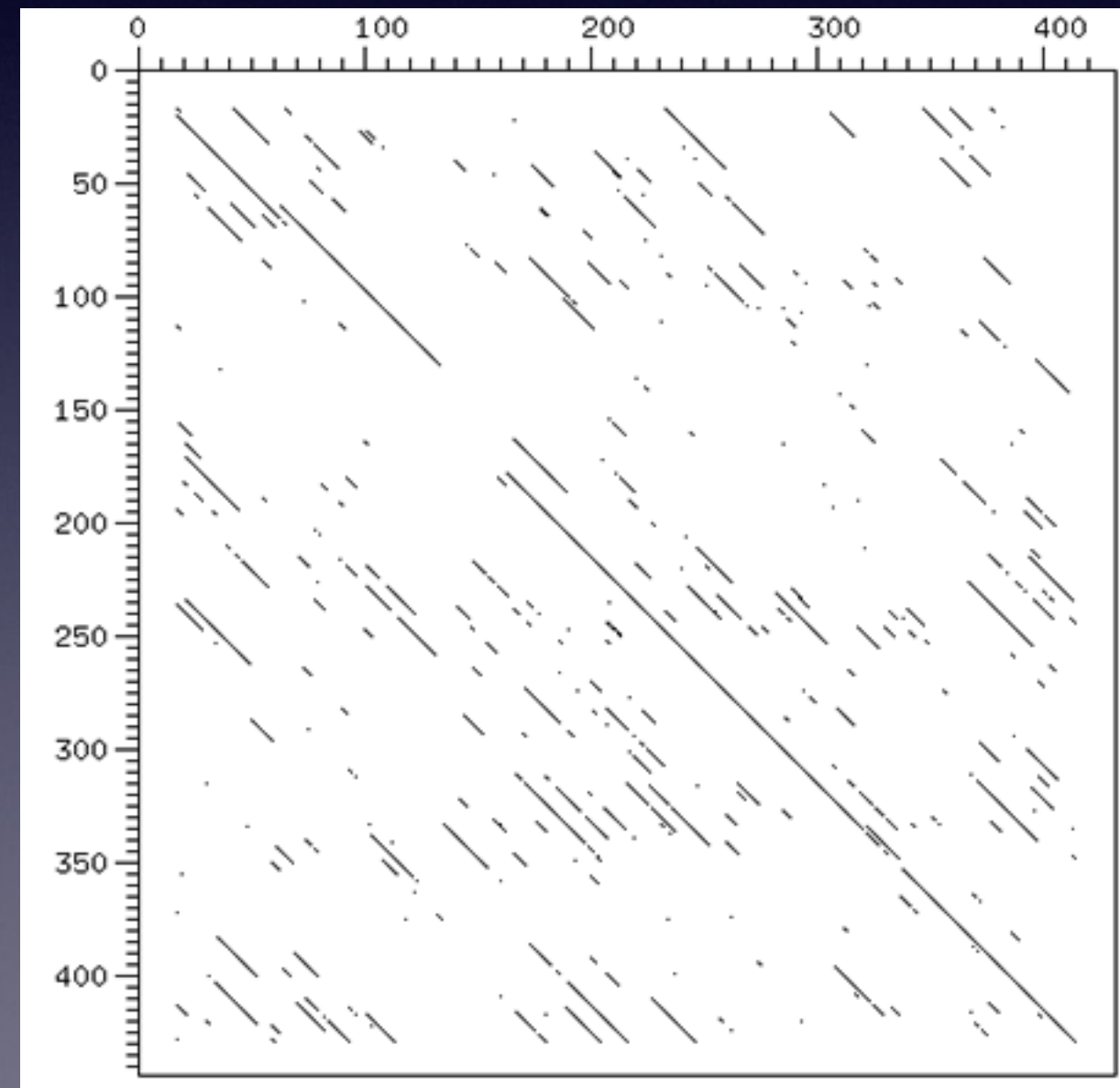
[illegible]

	A	C	K	F	L	F	G	D	E	L	R
C											
K											
F											
A											
R											
L											
F											
G											
D											
E											
L											

Realistic Dot Plot



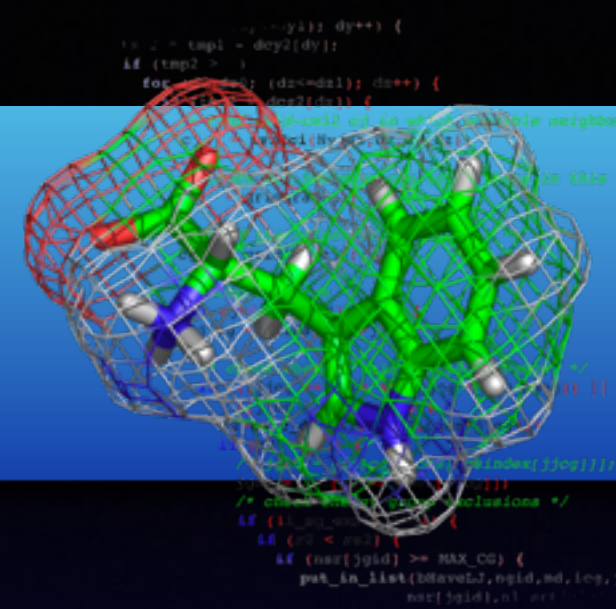
- Hemoglobin α chain vs. β chain
- Lots of false hits
- Hard to quantify



[illegible]

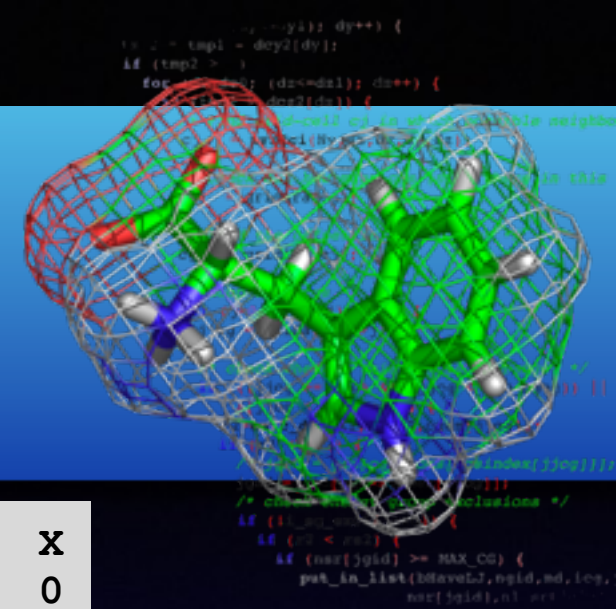
- What do we mean by “similar”?
 - Must it cover the whole sequence?
 - Do we allow gaps?
- Any way of pairing residues/gaps in the sequences is called an *alignment*
- Good alignments maximize similarity without adding too many gaps

Similarity Measures



- Amino acid substitution scores
 - Conserved amino acids (very good)
 - Similar amino acids (OK)
 - Neutral
 - Significantly different (very bad)
- Substitution scores: 20*20 matrix
- Example matrices: PAM250, BLOSUM62

BLOSUM62



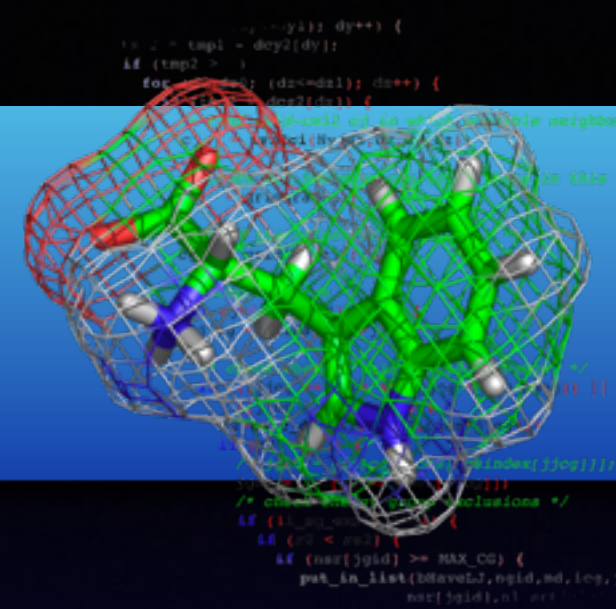
	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z	X
A	4	-1	-2	-2	0	-1	-1	0	-2	-1	-1	-1	-1	-2	-1	1	0	-3	-2	0	-2	-1	0
R	-1	5	0	-2	-3	1	0	-2	0	-3	-2	2	-1	-3	-2	-1	-1	-3	-2	-3	-1	0	-1
N	-2	0	6	1	-3	0	0	0	1	-3	-3	0	-2	-3	-2	1	0	-4	-2	-3	3	0	-1
D	-2	-2	1	6	-3	0	2	-1	-1	-3	-4	-1	-3	-3	-1	0	-1	-4	-3	-3	4	1	-1
C	0	-3	-3	-3	9	-3	-4	-3	-3	-1	-1	-3	-1	-2	-3	-1	-1	-2	-2	-1	-3	-3	-2
Q	-1	1	0	0	-3	5	2	-2	0	-3	-2	1	0	-3	-1	0	-1	-2	-1	-2	0	3	-1
E	-1	0	0	2	-4	2	5	-2	0	-3	-3	1	-2	-3	-1	0	-1	-3	-2	-2	1	4	-1
G	0	-2	0	-1	-3	-2	-2	6	-2	-4	-4	-2	-3	-3	-2	0	-2	-2	-3	-3	-1	-2	-1
H	-2	0	1	-1	-3	0	0	-2	8	-3	-3	-1	-2	-1	-2	-1	-2	-2	2	-3	0	0	-1
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4	2	-3	1	0	-3	-2	-1	-3	-1	3	-3	-3	-1
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4	-2	2	0	-3	-2	-1	-2	-1	1	-4	-3	-1
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5	-1	-3	-1	0	-1	-3	-2	-2	0	1	-1
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5	0	-2	-1	-1	-1	-1	1	-3	-1	-1
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6	-4	-2	-2	1	3	-1	-3	-3	-1
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7	-1	-1	-4	-3	-2	-2	-1	-2
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4	1	-3	-2	-2	0	0	0
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5	-2	-2	0	-1	-1	0
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11	2	-3	-4	-3	-2
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	-1	-3	-2	-1
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4	-3	-2	-1
B	-2	-1	3	4	-3	0	1	-1	0	-3	-4	0	-3	-3	-2	0	-1	-4	-3	-3	4	1	-1
Z	-1	0	0	1	-3	3	4	-2	0	-3	-3	1	-1	-3	-1	0	-1	-3	-2	-2	1	4	-1
X	0	-1	-1	-1	-2	-1	-1	-1	-1	-1	-1	-1	-1	-1	-2	0	0	-2	-1	-1	-1	-1	-1

B=D or N (Asp or Asn)
(Glu or Gln)

Z=E or Q

X=any amino acid

Alignment Scoring



- *We could* define any scoring we want
- Use a simple setup for two examples:
Match=3, Mismatch=-1, Gap=-2

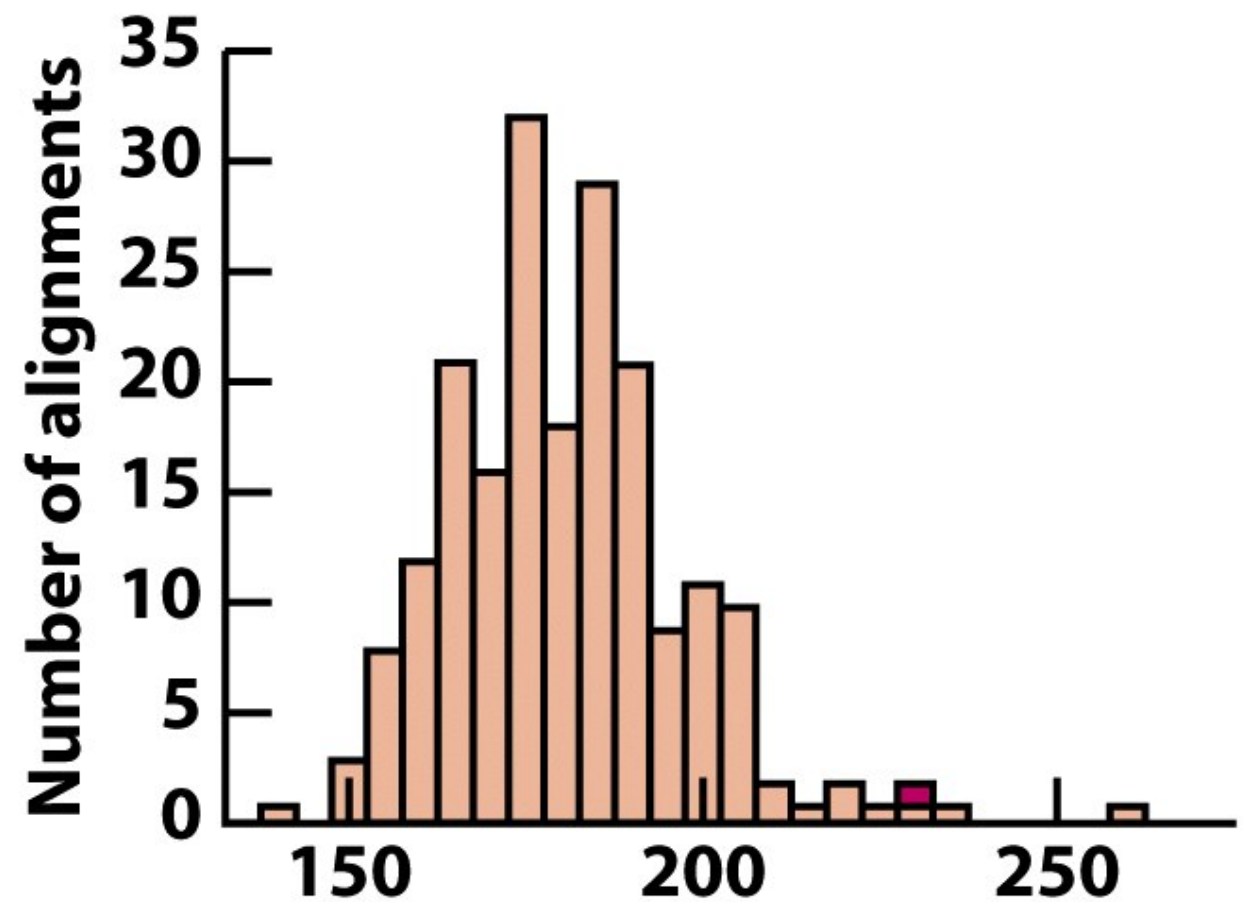
1 DEFYWLKKPAGTSVQND
 | | | | | | | | |
 EEFYWKKPAGTSAVQND

Score: 19

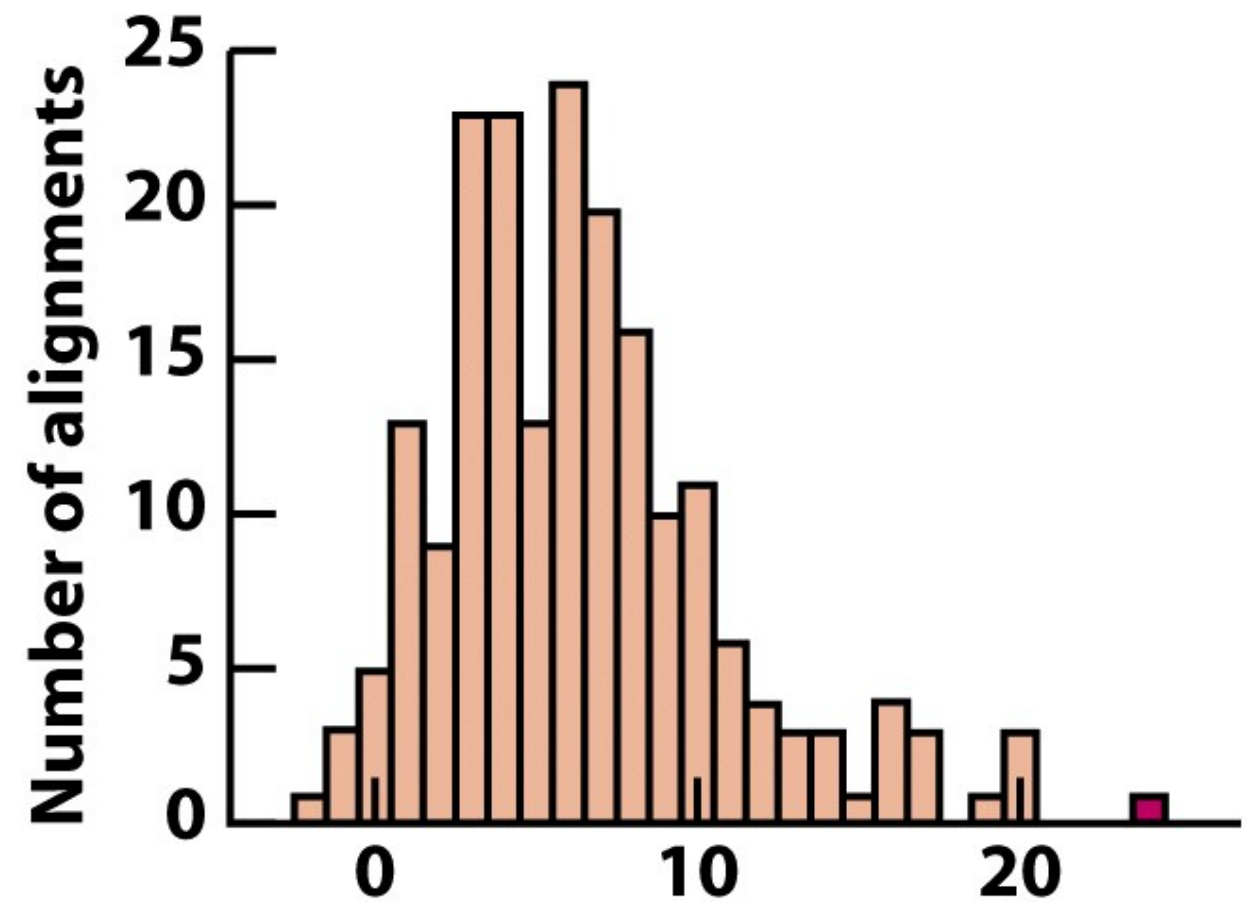
2 DEFYWLKKPAGTS-VQND
 | | | | | | | | | | | | | |
 EEFYW-KKPAGTSAVQND

Score: ^{Better!} 40

Similarity better than identity for alignments!



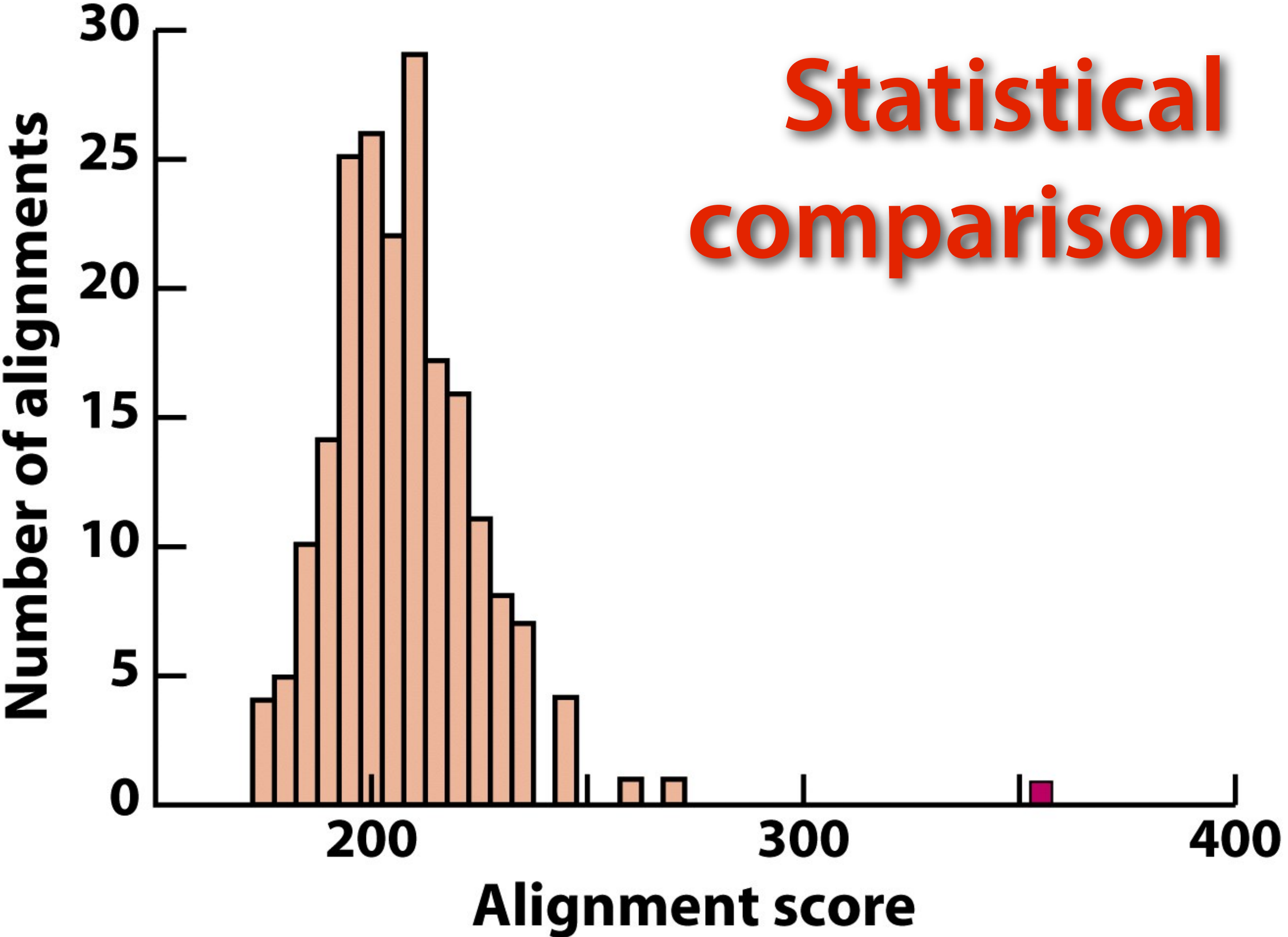
(A) Alignment score (identities only)



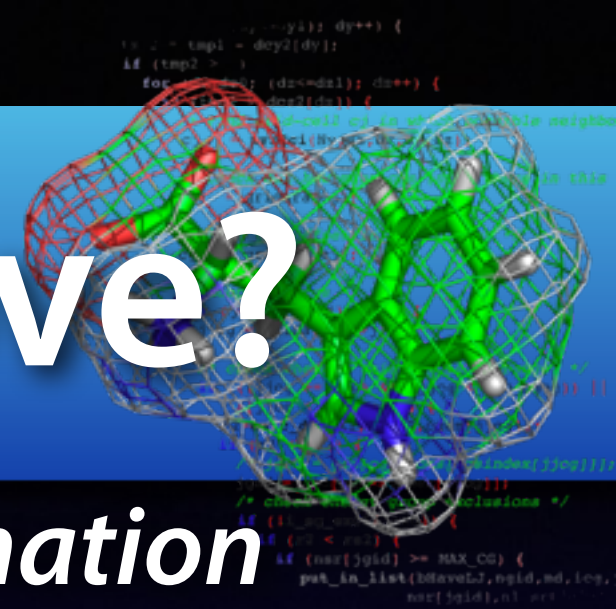
(B) Alignment score (Blosum 62)

Figure 6-11
Biochemistry, Sixth Edition
© 2007 W. H. Freeman and Company

© 2007 W. H. Freeman and Company
Biochemistry, Sixth Edition
Figure 6-11



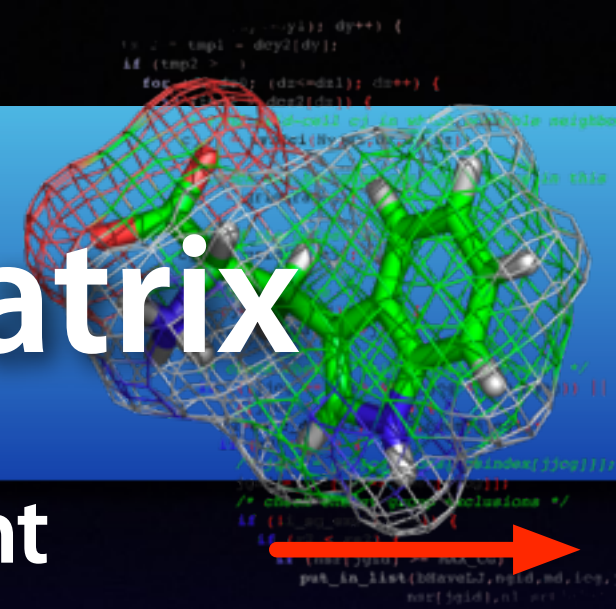
How can we improve?



- The key here was *evolutionary information*
- Can you find and use more such data?

MYHU	..MGLSDGEWQLVLNVWGKVEADIPGHGQEV LIRLFKGHPE..
MYCZ	...GLSDGEWQLVLNVWGKVEADIPGHGQEV LIRLFKGHPE..
MYMQV	...GLSDGEWQLVLNIWGKVEADIPSHGQEV LISLFKGHPE..
MYOY	...GLSDAEWQLVLNVWGKVEADIPGHGQDVLIRLFKGHPE..
MYFXBE	...GLSDGEWQIVLNIWGKVETDLAGHGQEV LIRLFKKNHPE..
MYDG	...GLSDGEWQIVLNIWGKVETDLAGHGQEV LIRLFKKNHPE..
MYWHL	...GLSDGEWQLVLNVWGKVEADLAGHGQDILIRLFKGHPE..
MYPN	...GLNDQEWQQVLTMWGKVESDLAGHGHAVLMR LFKSHPE..
MYTUYADFDAVLKCVGPVEADYT T MGGLVLT RLFKEHPE..
Consensus	GLSDG ewQL N K A GH QEv IR G

Position-Specific Scoring Matrix



Position in our multiple sequence alignment

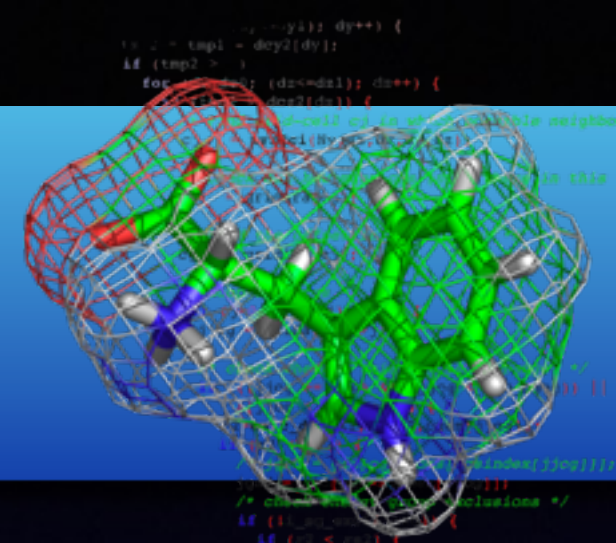
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21

Amino acids

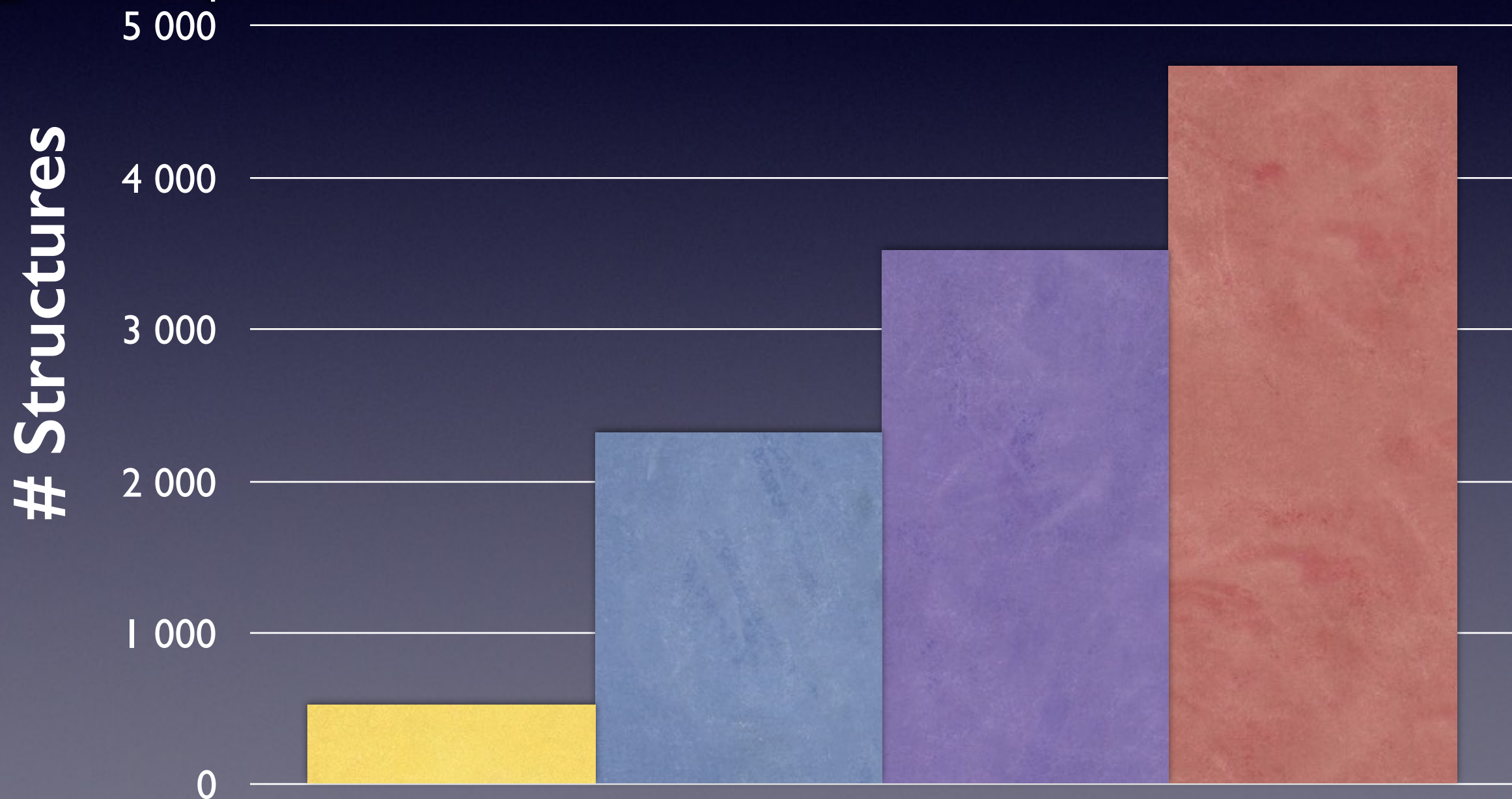
A
C
D
E
F
G
H
I
K
L
M
N
P
Q
R
S
T
V
W
Y

-3	3	0	1	2	1	0	1	-4	2	0	-1	2	-1	-3	-4	2	2	1	1	2
4	2	-4	0	1	2	1	6	2	-1	1	2	4	-4	3	-2	-1	4	0	6	4
0	-4	-4	-2	-4	1	-2	3	2	3	3	-3	5	2	1	1	3	5	-2	3	5
-2	3	2	-3	1	0	-4	2	-3	2	-1	3	-1	0	3	0	2	-1	-3	2	-1
1	-2	0	-5	1	0	3	-2	3	3	3	0	-2	2	1	1	3	-2	-5	-2	-2
3	3	3	1	0	-2	0	-1	-1	4	-4	-4	7	0	2	0	4	7	1	-1	7
0	0	-2	-3	4	-4	-4	5	2	5	4	0	-6	2	1	1	5	-6	-3	5	-6
-4	1	0	1	2	0	4	3	-1	2	0	-1	4	1	3	2	2	4	1	3	4
3	-4	-1	3	0	0	3	2	-4	-4	-5	3	3	-2	-1	1	-4	3	3	2	3
3	1	-4	0	4	0	4	1	3	-1	0	-4	-3	-5	4	1	-1	-3	0	1	-3
2	-1	1	2	2	1	-1	3	3	3	2	-1	0	2	0	4	3	0	2	3	0
1	2	2	-3	-3	2	1	2	0	-1	-4	-2	0	-3	3	1	-1	0	-3	2	0
-4	-5	2	3	3	-2	3	3	0	2	4	1	-1	-2	3	-2	2	-1	3	3	-1
3	2	5	4	0	3	0	5	5	-4	5	0	2	0	-1	-1	-4	2	4	5	2
1	-4	1	-2	-4	0	0	4	-1	3	1	-2	0	2	4	5	3	0	-2	4	0
2	0	0	2	-2	4	2	1	-1	0	4	-4	-1	-4	2	4	0	-1	2	1	-1
1	1	-4	4	0	1	-1	3	1	-4	2	-2	0	-1	-1	-4	-4	0	4	3	0
-2	4	2	2	1	5	1	-4	-1	1	1	2	2	0	2	3	1	2	2	-4	2
0	5	-4	0	3	3	3	2	3	2	4	3	-1	1	2	-4	2	-1	0	2	-1
5	-2	1	-2	2	4	-1	2	1	3	3	0	-2	-1	1	3	3	-2	-2	2	-2

Search sensitivity



■ Predictions with sequence ■ Predictions with profile ■ Predictions with HMM
■ Total sequences

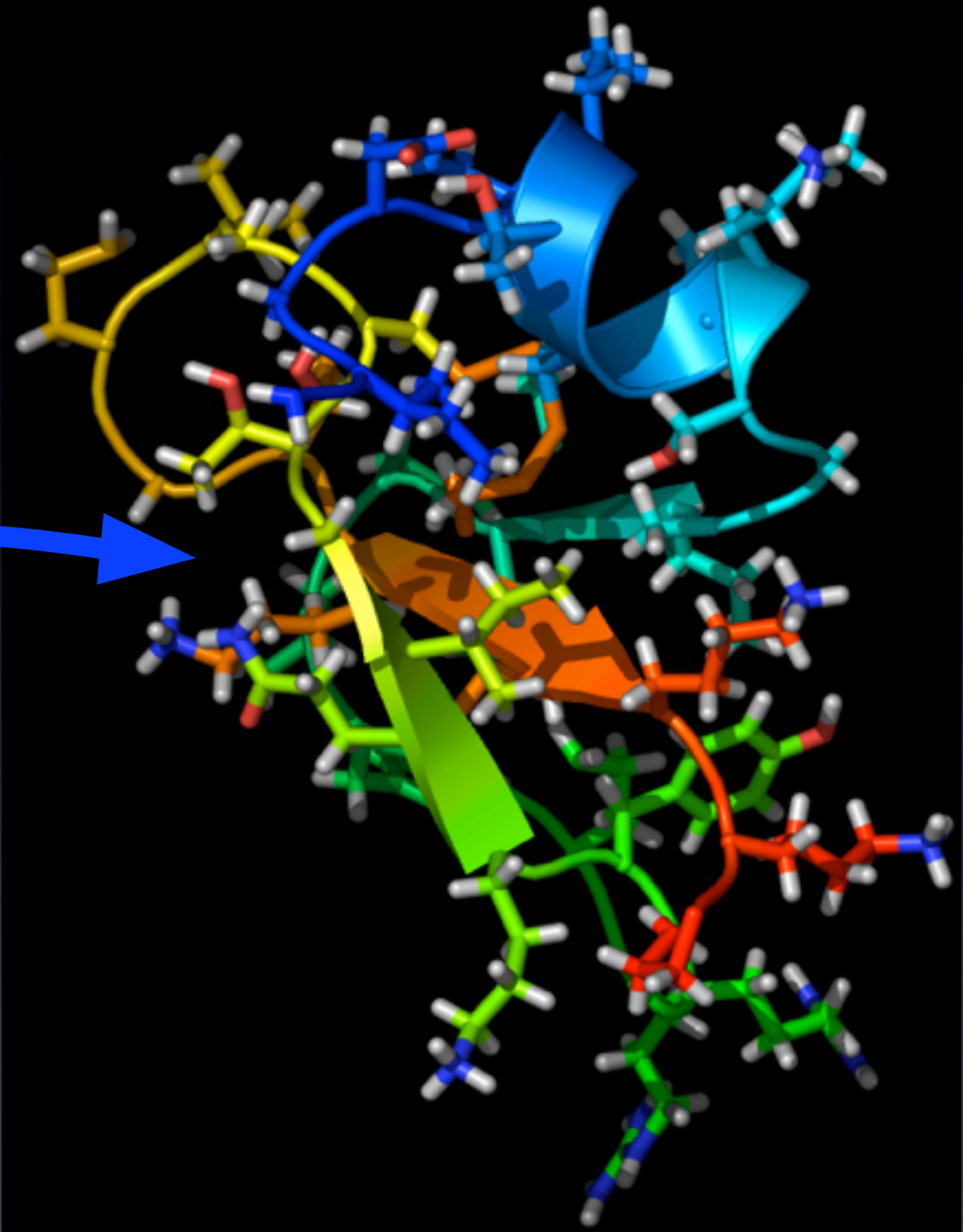


Proteins in the Shewanella Oneidensis genome

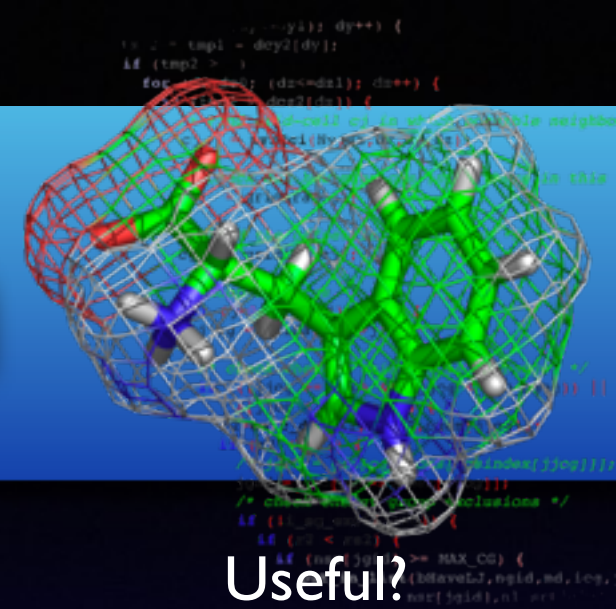
Protein Structure Classification & Prediction

KIEEGKLVIW	INGDKGYNGL
AEVGKKFEKD	TGIKVTVEHP
DKLEEKFPQV	AATGDGPDII
FWAHDREGGY	AQSGLLAEIT
PDKAFQDKLY	PFTWDAVRYN
GKLIAYPIAV	EALSLIYNKD
LLPNPPKTWE	EIPALDKELK
AKGKSALMFN	LQEPYFTWPL
IAADGGYAFK	YENGKYDIKD
VGVDNAGAKA	GLTFLVDLIK
NKHMNADTDY	SIAEAAFNKG
ETAMTINGPW	AWSNIDTSKV
NYGVTVLPTF	KGQPSKPFVG
VLSAGINAAS	PNKELAKEFL
ENYLLTDEGL	EAVNKDKPLG
AVALKSYEEE	LAKDPRIAAT
MENAQKGEIM	PNIPQMSAFW
YAVRTAVINA	ASGRQTVDEA

Not quite trivial...



Structure prediction



Method

Knowledge

Approach

Difficulty

Useful?

Secondary
structure
prediction

Sequence-structure
statistics

Predict helix, strand, or coil
for each residue

Medium

Sometimes
(membrane prots.)

Homology
modeling

Homologs of known
structure

Identify sequence homologs,
copy 3D coords and modify

Fairly easy

Quite reliable with
high identity. Use in
drug design.

Fold
recognition

Proteins of known
structure

Assemble parts from
(several) proteins - often not
homologs

Medium to
hard

More of a long
shot, but models
are often correct

Ab initio

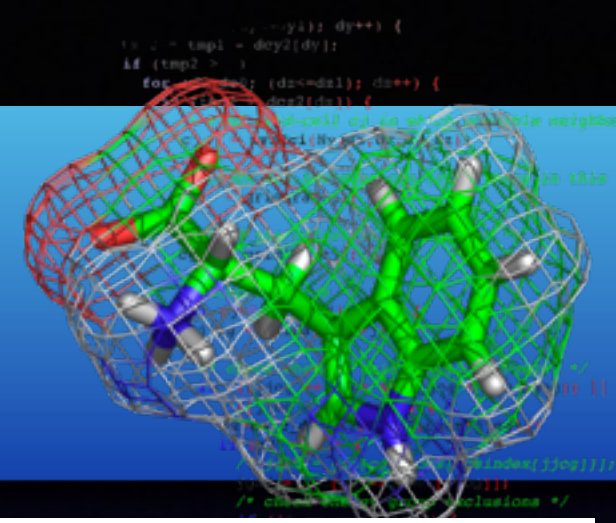
Physics and general
biology statistics

Simulate folding, or generate
lots of structures and pick
the best one

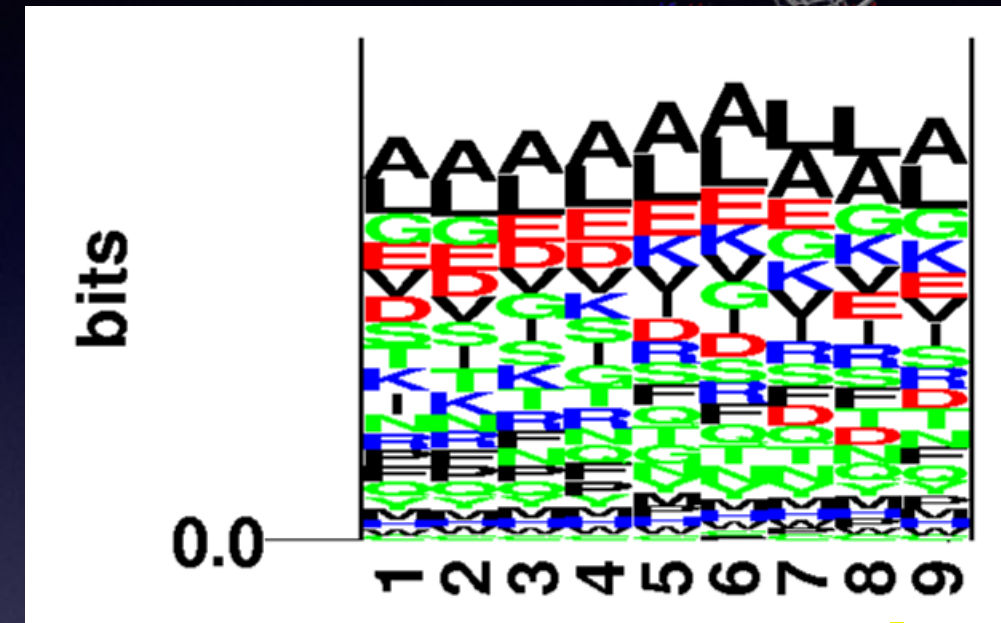
Extremely
hard

Does not yet work
reliably. Too hard?

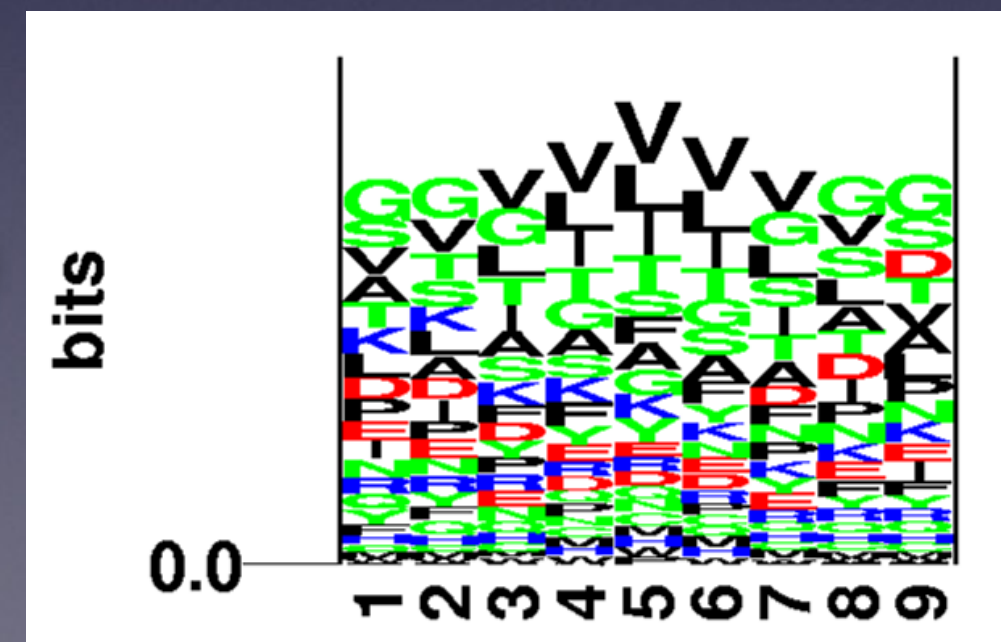
Secondary structure



- Hydrophobicity patterns in helices/strands
- AA Preferences for helix/strand/coil
- Best methods reach ~80% accuracy
- Special case: Predicting transmembrane helices and their in/out topology!



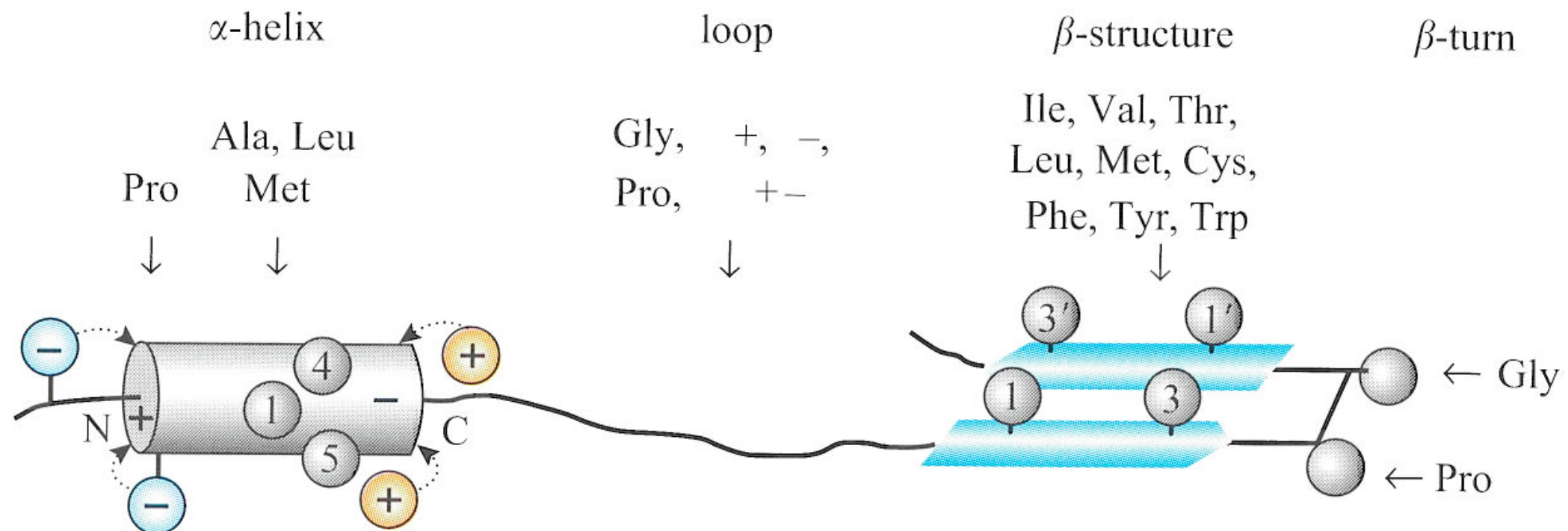
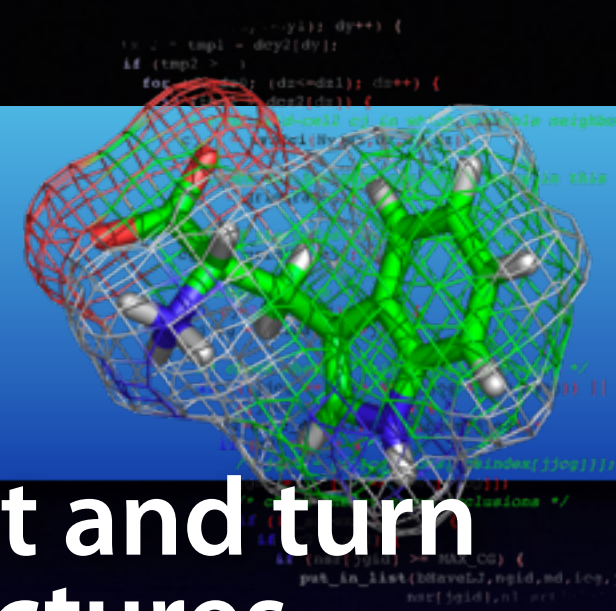
α -Helix



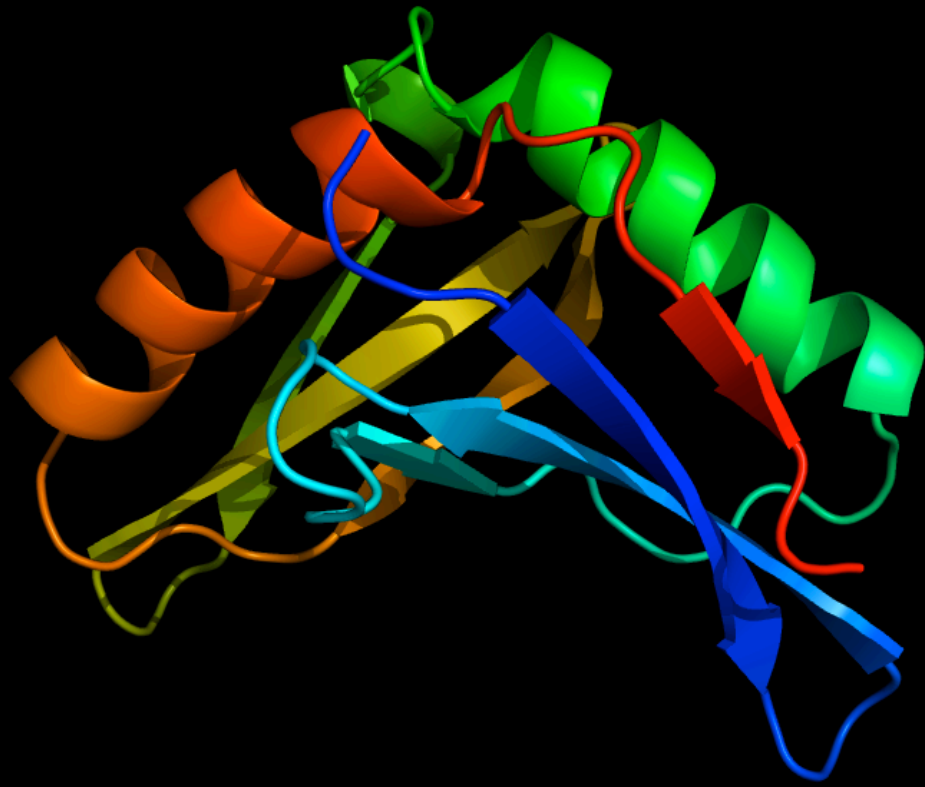
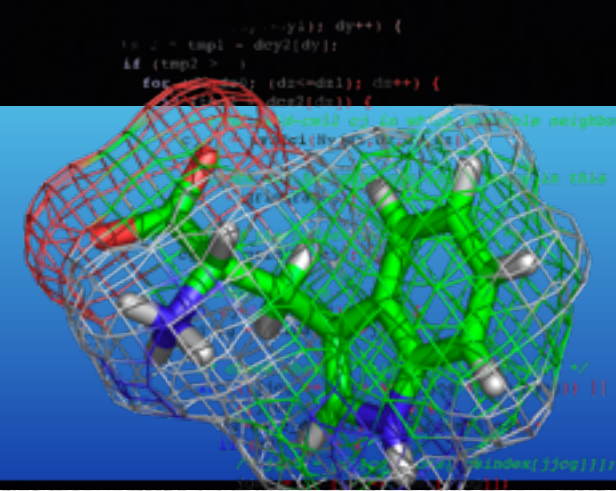
β -Strand

Chou-Fasman

- Determine the probability of helix, sheet and turn for each residue based on available structures
- Single unfavorable residues can occur
- But the rolling average properties of amino acids should be a useful predictor



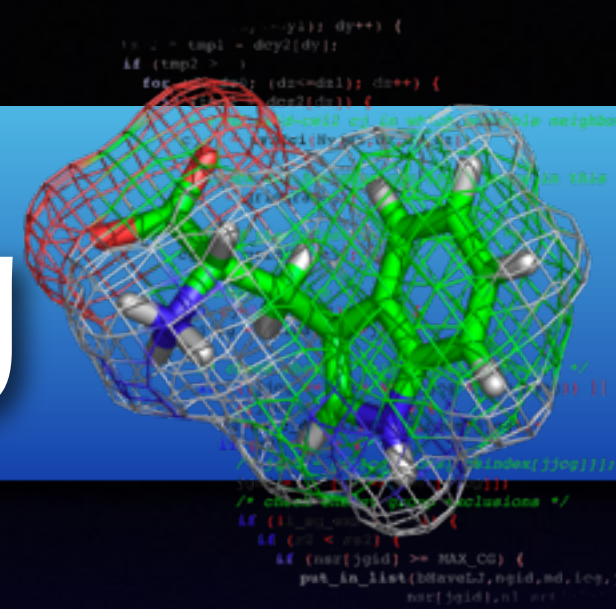
Chou-Fasman data



“Propensity” rather than probability, but it contains the same information

Name	P(a)	P(b)	P(turn)	f(i)	f(i+1)	f(i+2)	f(i+3)
Alanine	142	83	66	0.060	0.076	0.035	0.058
Arginine	98	93	95	0.070	0.106	0.099	0.085
Aspartic acid	101	54	146	0.147	0.110	0.179	0.081
Asparagine	67	89	156	0.161	0.083	0.191	0.091
Cysteine	70	119	119	0.149	0.050	0.117	0.128
Glumatic acid	151	37	74	0.056	0.060	0.077	0.064
Glutamine	111	110	98	0.074	0.098	0.037	0.098
Glycine	57	75	156	0.102	0.085	0.190	0.152
Histidine	100	87	95	0.140	0.047	0.093	0.054
Isoleucine	108	160	47	0.043	0.034	0.013	0.056
Leucine	121	130	59	0.061	0.025	0.036	0.070
Lysine	114	74	101	0.055	0.115	0.072	0.095
Methionine	145	105	60	0.068	0.082	0.014	0.055

Homology modeling

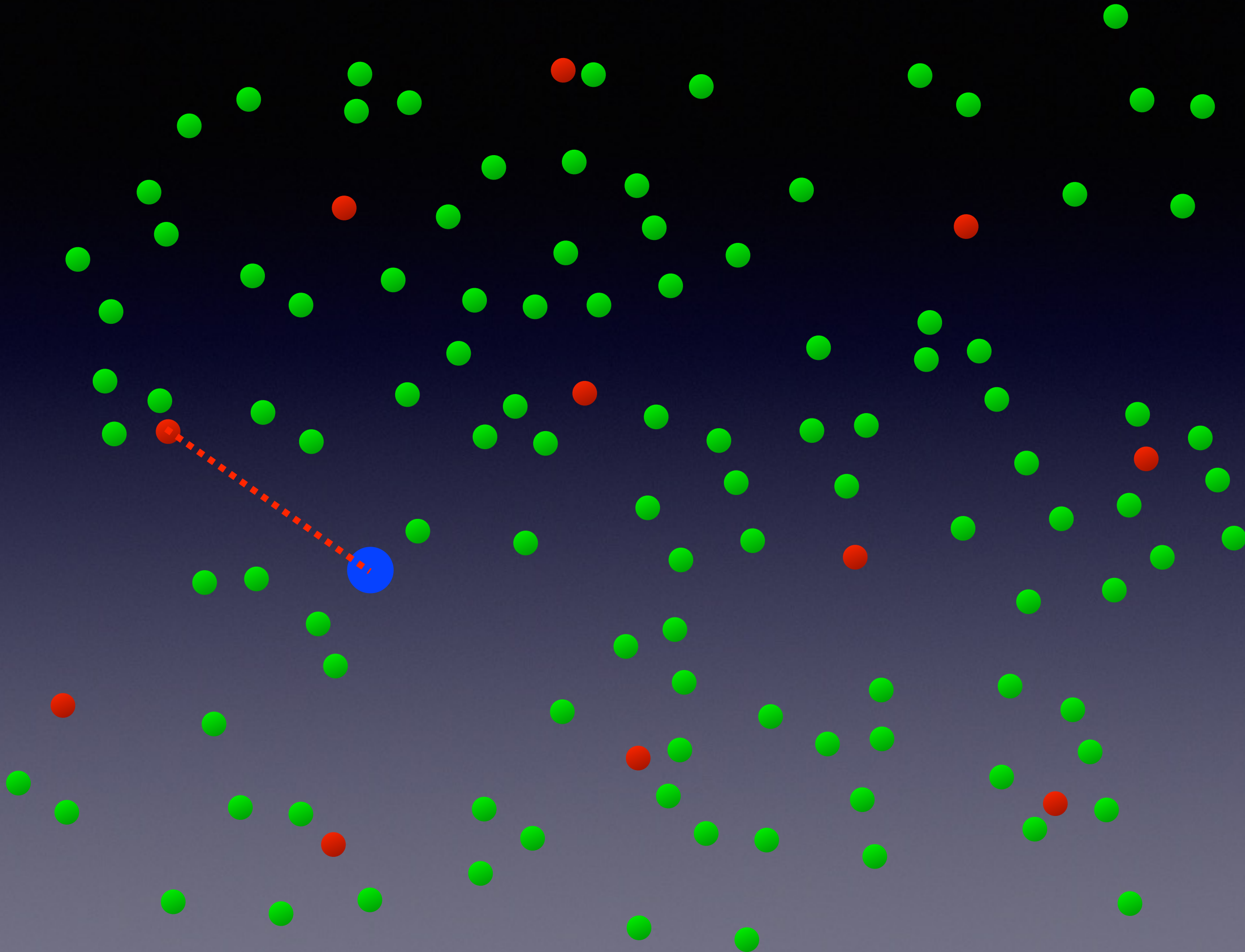


- Protein structures are stable
 - Small sequence changes usually only lead to small variations in 3D structure
- Insertions/deletions usually occur in loop regions, not in helices or sheets
- Sequence matching methods are very good at finding homologs
- Ideally you only need to rebuild side chains

Model Quality?



- Depends on modeling distance
- 95% identical residues: perfect model
- 20% identical residues: questionable
- Structural Genomics
 - Reducing modeling distances by determining more 3D crystal structures



**We only need experimental
structures for a set of
representative folds to
create reasonable models
for 90% of proteins**

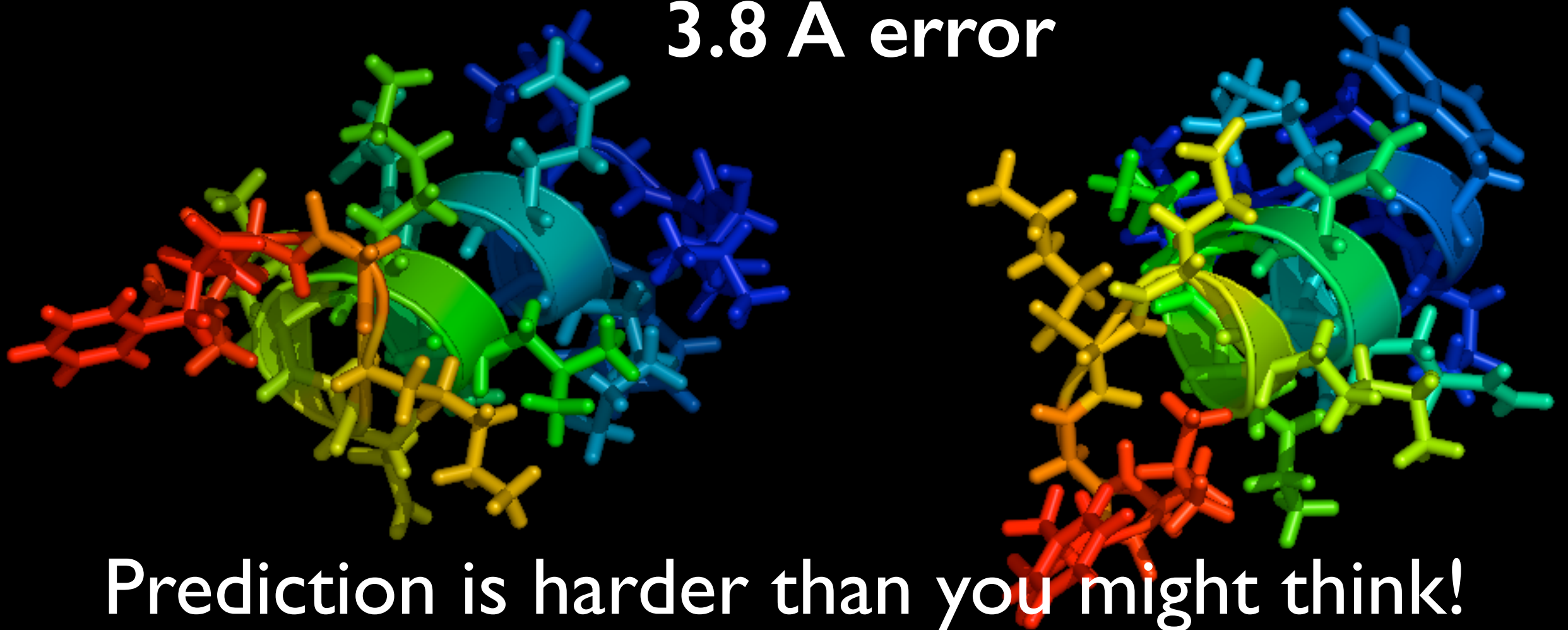
**Goal of the Structural
Genomics Project is 100,000
new structures**

The Alignment Problem

Template FVNQHLCGSHLVEALYLVCGERGFFCCTSICSLYQ

Query FYTFKGIVEQCCTSICSLYQLENYCNQHLCGSHLV

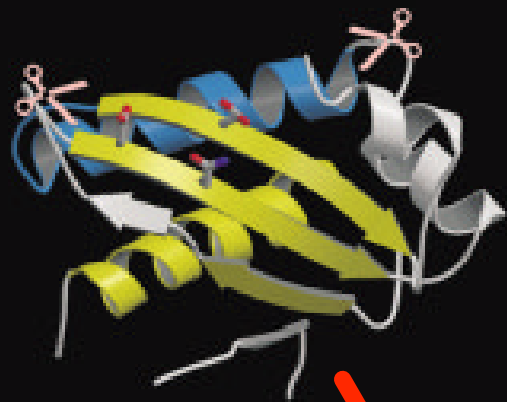
3.8 Å error



Prediction is harder than you might think!

Multiple Templates

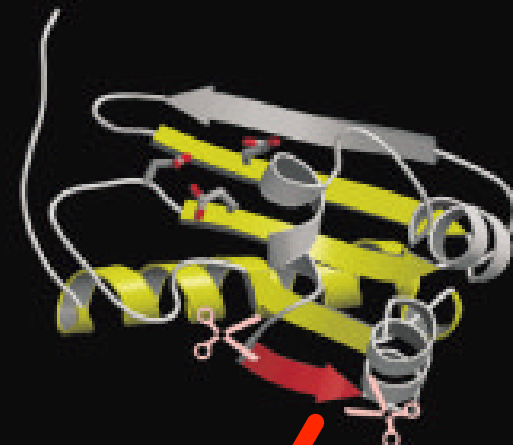
ERA GTPase



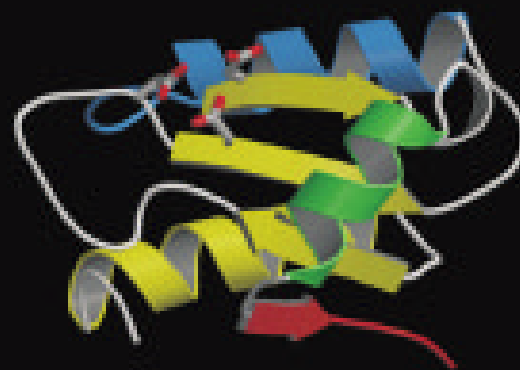
Poly(A) polymerase



Kanamycin
nucleotidyltransferase

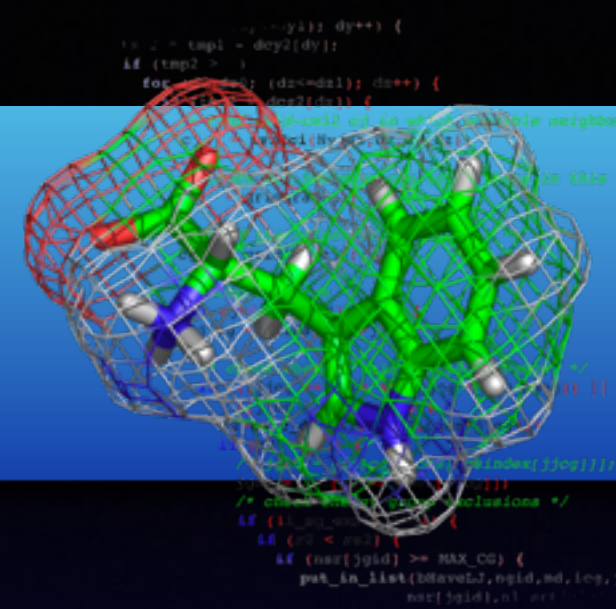


H10073, *H. Influenzae* (T0130)



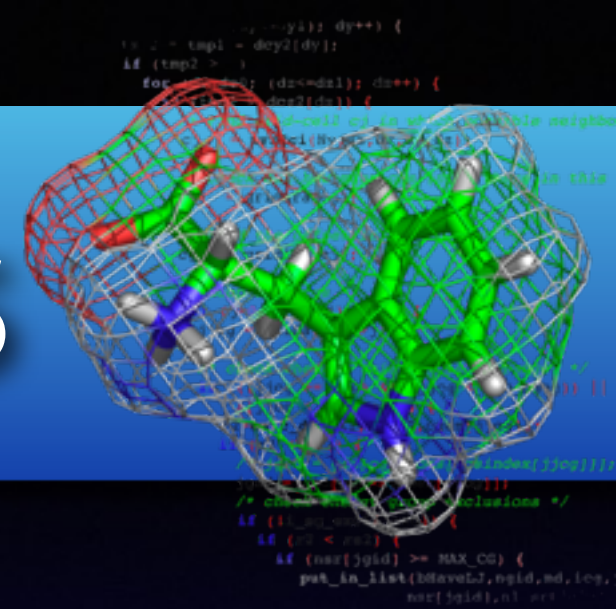
Conserved core,
combined with
different elements

Ab Initio prediction



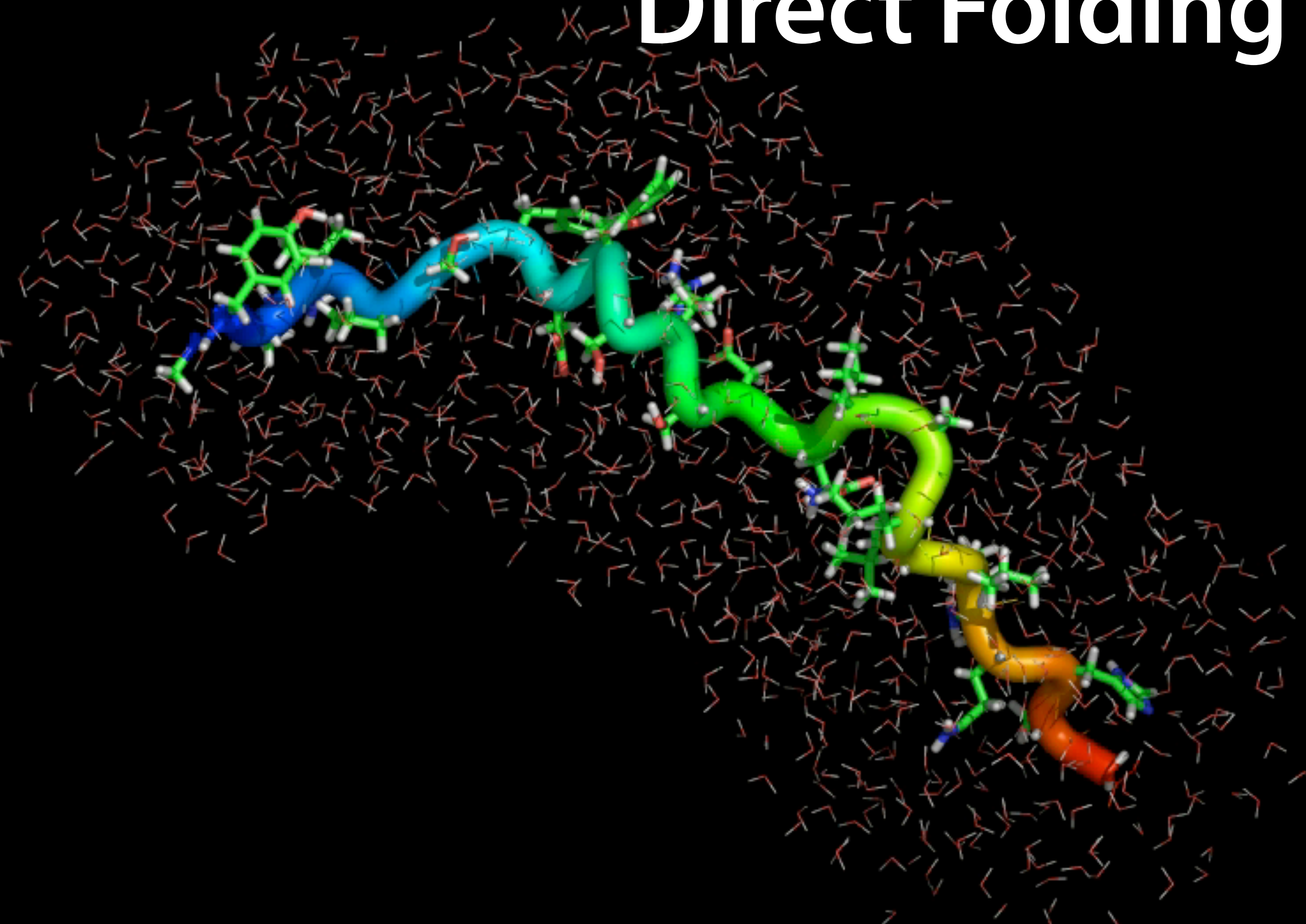
- Consider a 100 residue protein
- Assume there are 10 conformations/aa
- 10^{100} structures to test
- Levintal's paradox: It would take the age of the universe to test everything
- In practice it must be a guided process
- But how do you do it **in a computer?**

Possible approaches

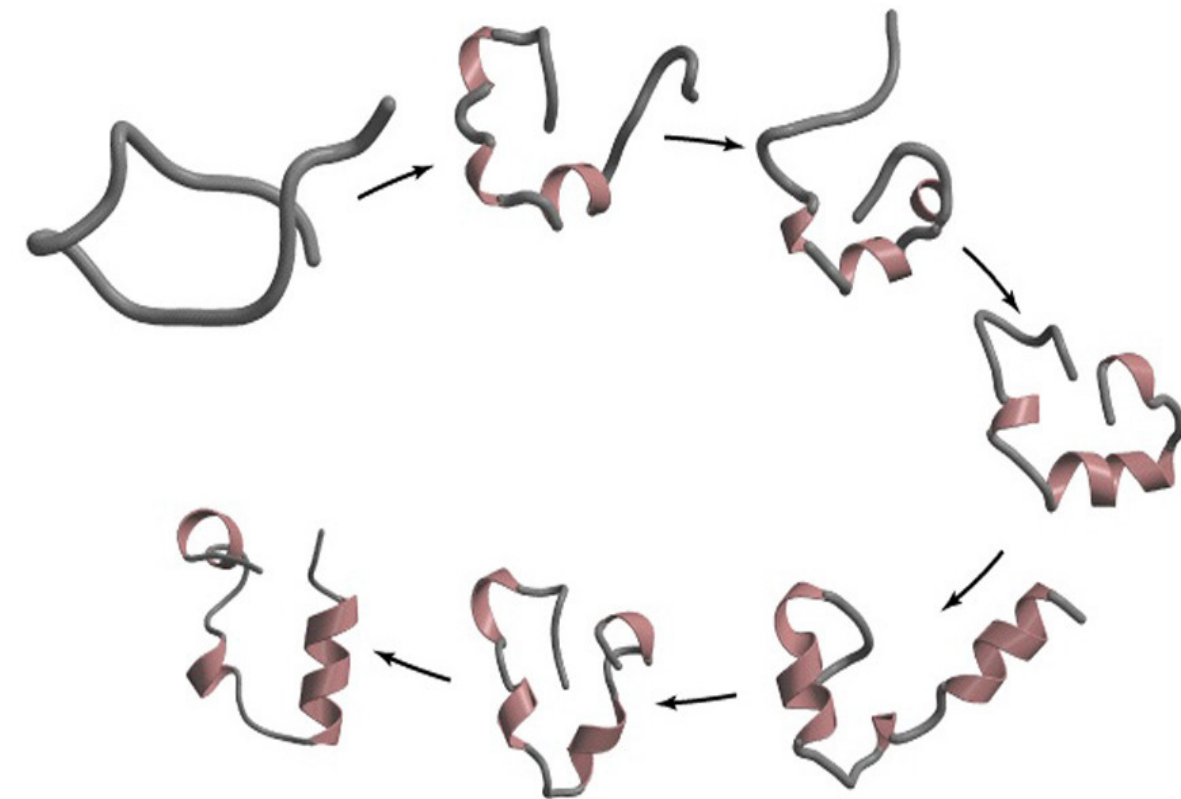
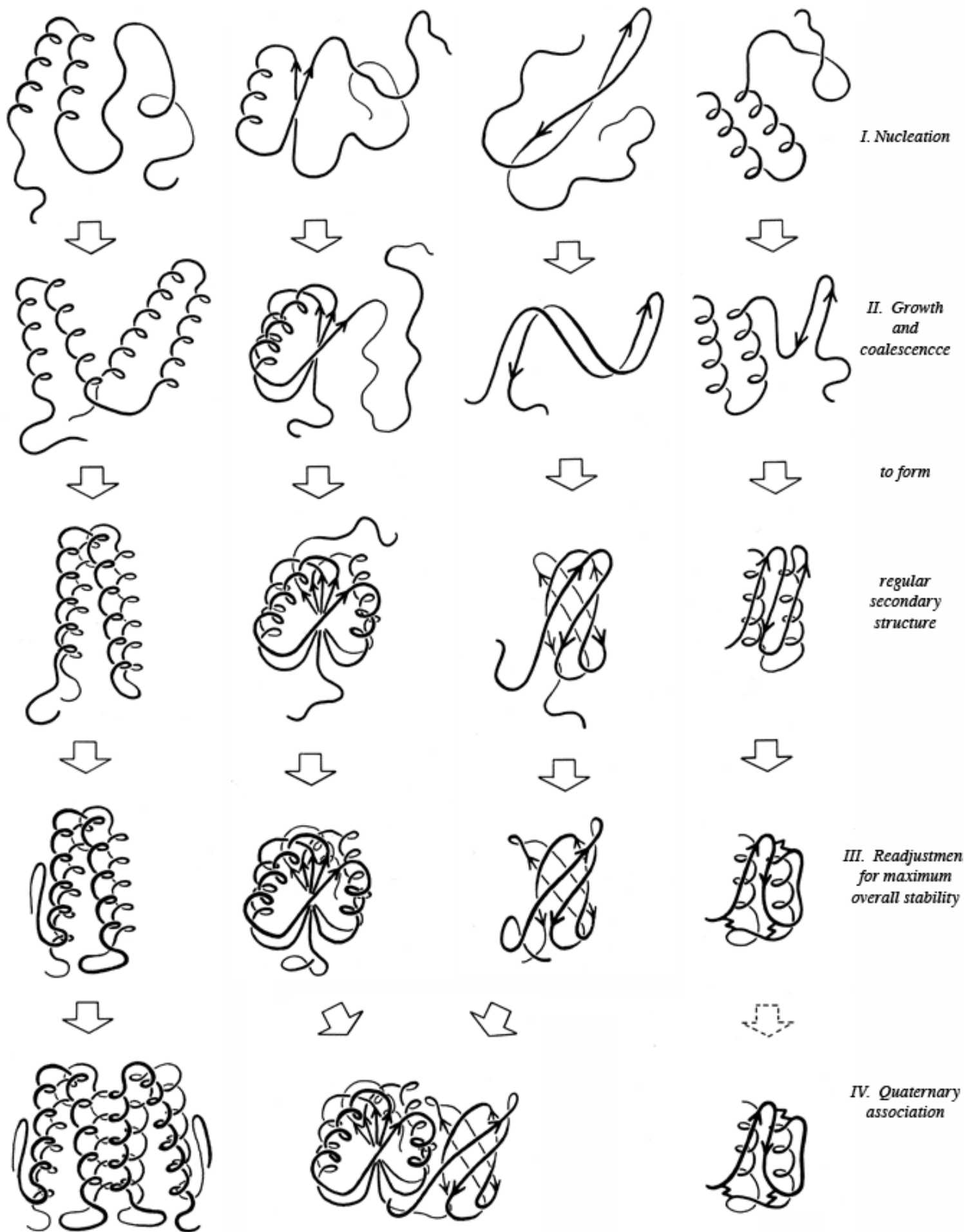


- Brute force physical simulation
 - Would provide both the path & goal
 - Even supercomputers are usually too slow
- Smarter *ab initio* algorithms
 - The path is usually NOT the goal
 - Create test structures & find the best
 - Fragment assembly: ROSETTA (Baker)

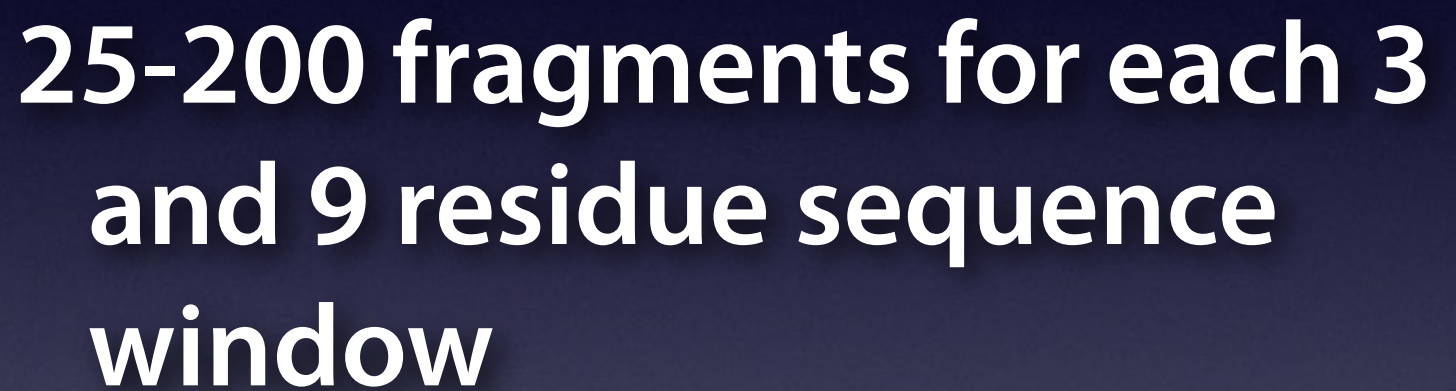
Direct Folding



The Rosetta idea

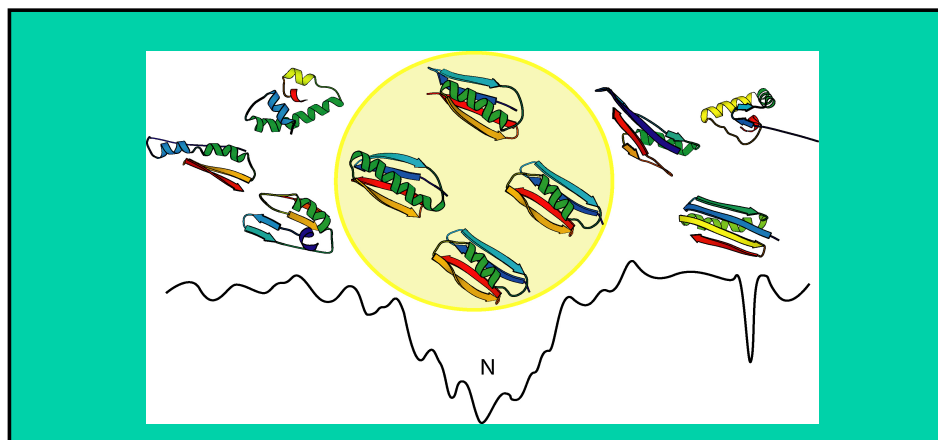
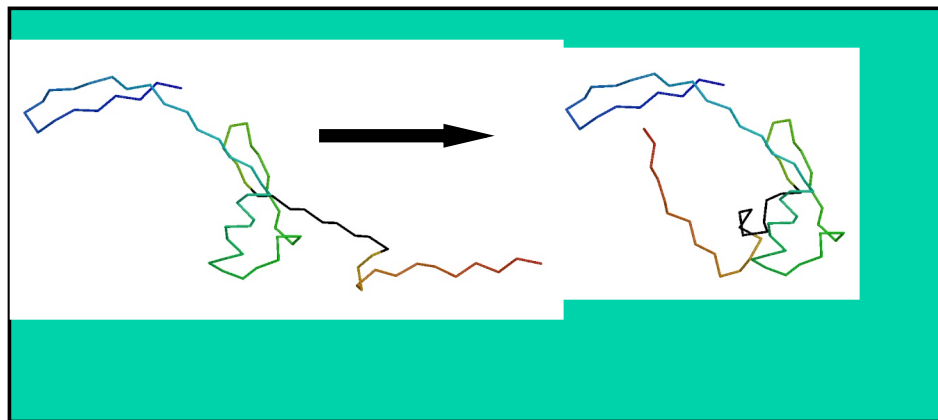
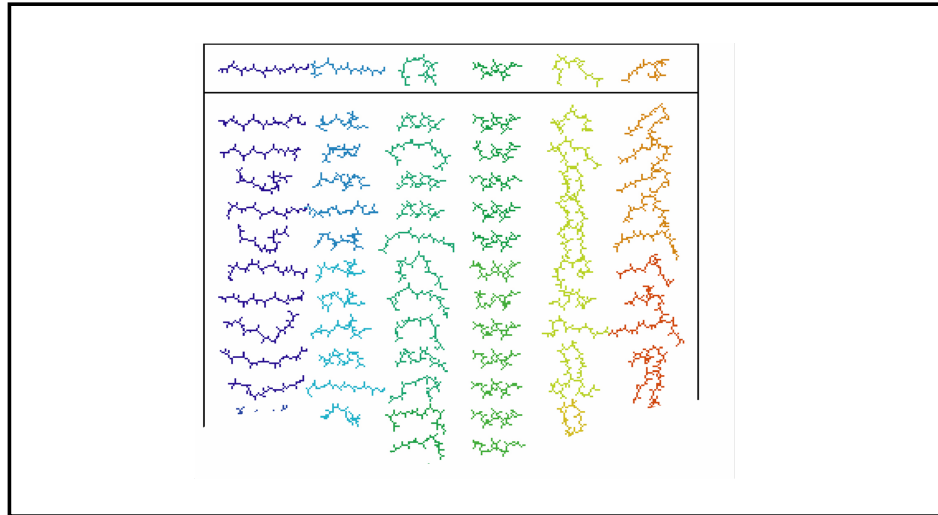


A 3D molecular model of a protein-ligand complex. The protein is shown as a green mesh surface with a red wireframe. The ligand is shown as a green stick model. A large white 'V' is overlaid on the model. The background is a blue gradient.



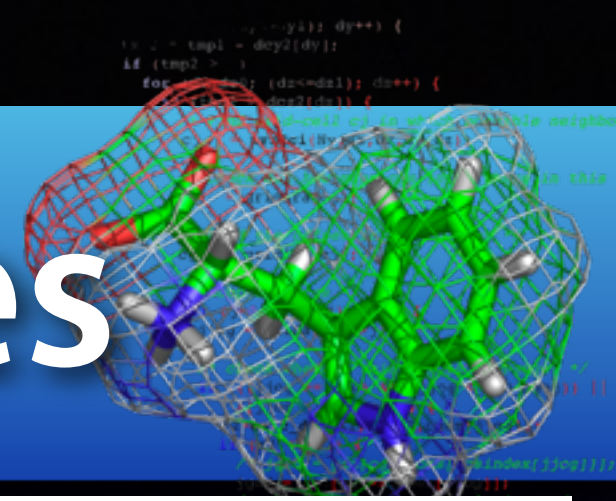
**Selected from known structures
Better than 2.5Å resolution
< 50% sequence identity**

Prediction with Rosetta

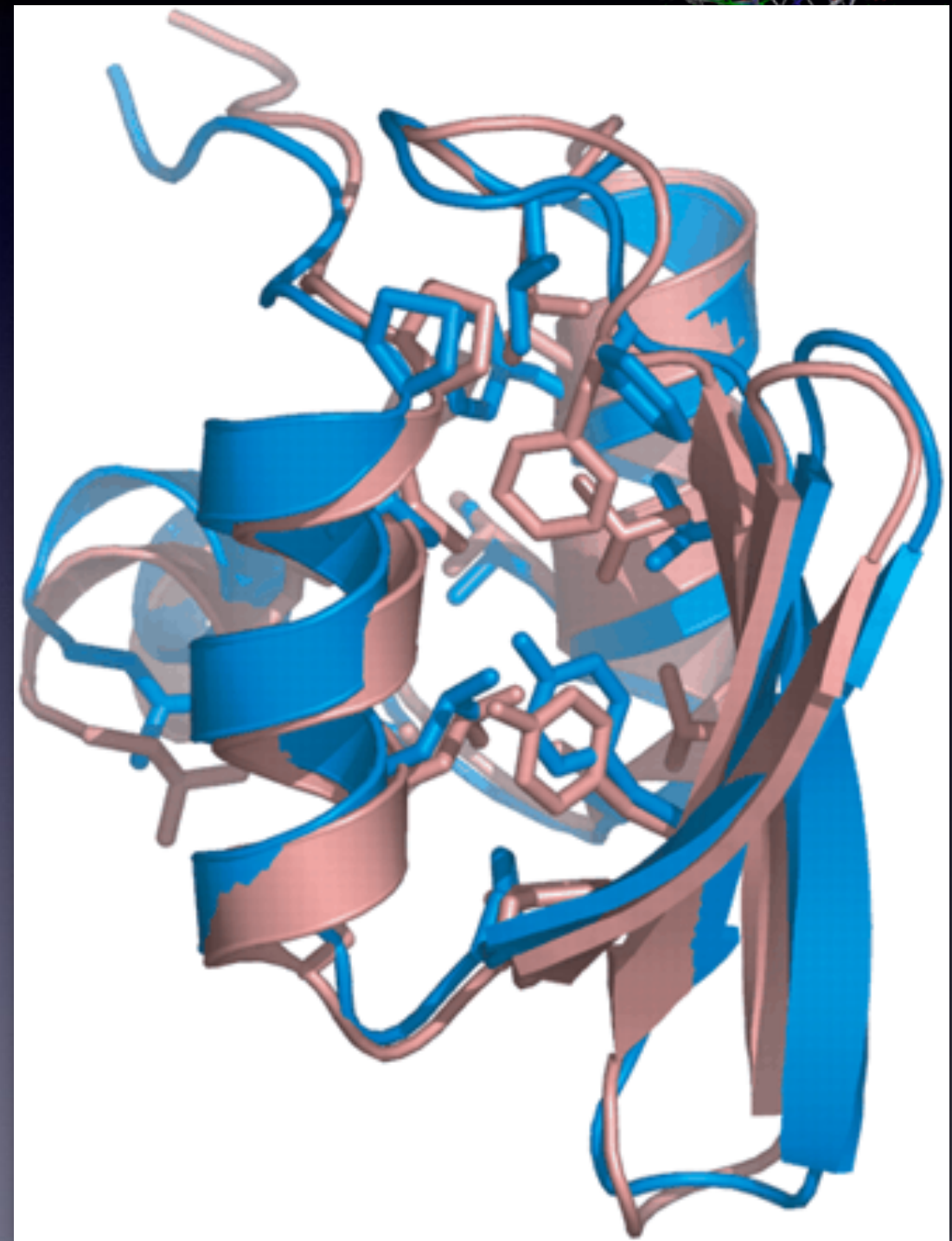


- Select fragments with good local properties
- Assemble into protein-like folds (lots of them)
- Use physics-based energy functions to try and select the best one

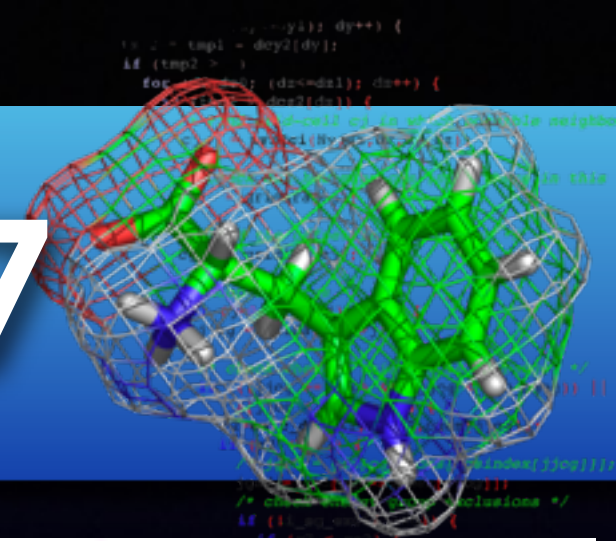
Rosetta Successes



- Refinement: Make small moves in torsion angles
- Rebuild sidechains
- Minimize energy
- Repeat refinement, etc.
- Bradley, Science 2005:
5 of 16 structures predicted
to within 1.5Å resolution!



Rosetta Design: TOP7



- Can you design a completely new fold not seen in nature?
- Iterate design & refinement
- Extremely stable structure
- Determined structure in experiments to confirm:
Less than 1.2Å difference!



INFORMATION + PHYSICS = LIFE

DNA Sequence → RNA Sequence → Protein Sequence → Folded Protein

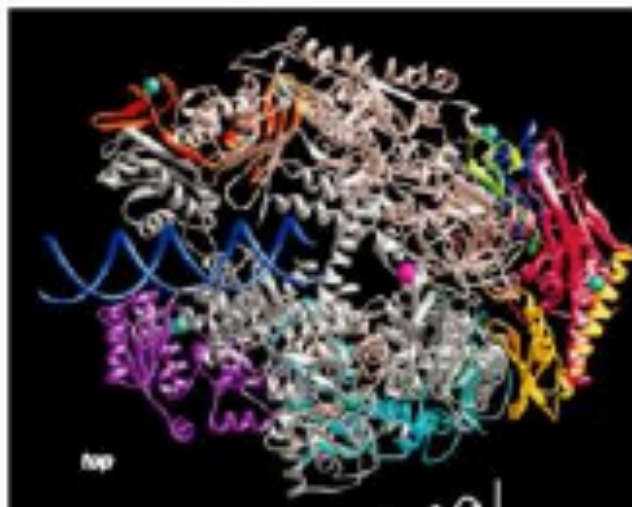
• in silico

Easy:
Change
T to U

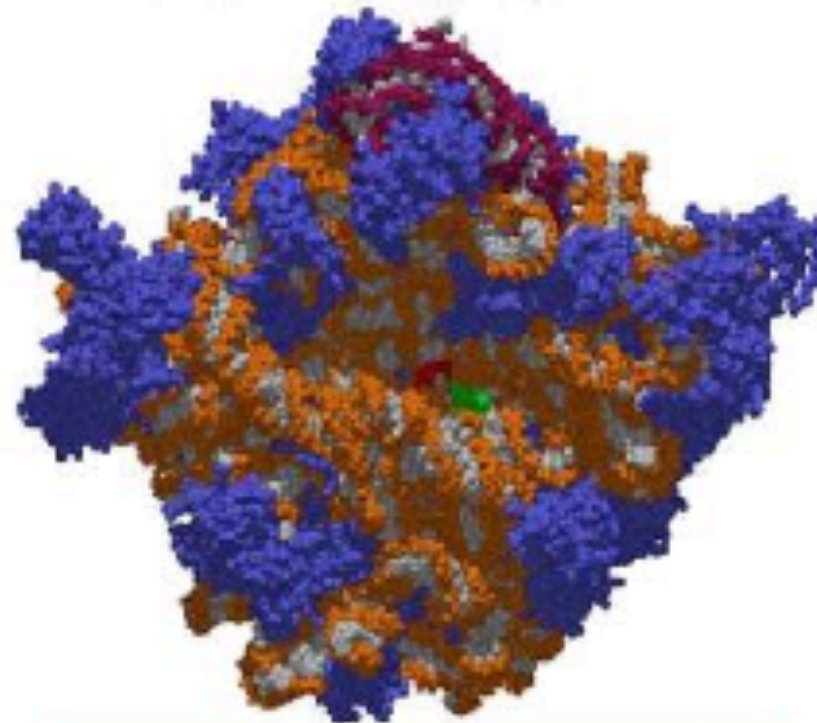
Easy:
Triplet
Code

Hard: Folding is
many body
simulation

• in vivo



Hard: Transcription
Polymerase



Hard: Translation
Ribosome

Easy: Folding is
free by laws of
physics

Michael Levitt