*Supplementary Information*

Fast search algorithms for Computational Protein Design

Seydou Traoré1-3, Kyle E. Roberts4, David Allouche5, Bruce R. Donald4, Isabelle André1-3, Thomas Schiex5, Sophie Barbe1-3\*

[[1]](#endnote-1) Université de Toulouse; INSA,UPS,INP; LISBP, 135 Avenue de Rangueil, F-31077 Toulouse, France

2 INRA, UMR792, Ingénierie des Systèmes Biologiques et des Procédés, F-31400 Toulouse, France

3 CNRS, UMR5504, F-31400 Toulouse, France

4 Department of Biochemistry, Department of Computer Science, Duke University, Durham, NC, USA

5 Unité de Mathématiques et Informatique Appliquées de Toulouse, UR 875, INRA, F-31320 Castanet Tolosan, France

\* Correspondence to: Sophie Barbe, Laboratoire d'Ingénierie des Systèmes Biologiques et des Procédés – INSA; CNRS UMR5504; UMR INRA 792; 135, Avenue de Rangueil; F-31077 Toulouse cedex 4, France. Tel: +33 561 559 963; Fax: +33 561 559 400; E-mail: sophie.barbe@insa-toulouse.fr

1. |  |
| --- |
| Table S-1 Benchmark set. For each instance: system reference PDB id (PDB), the number of protein residues (N), the number of designable residues of the whole system (N-designable is the sum of mutable residues for which amino acid identity changes are allowed and flexible residues for which only conformational side-chain flexibility is allowed), the minimal and maximal number of rotamers per variable position, the size of the total search space size [1] and the time for computing the energy matrix. |
| *PDB* | *N* | *N-designable (mutable, flexible)* | *N-rotamer range min-max per residue* | *Sequence-Conformation Space Size* | *Energy Matrix Computation Time (sec)* |
| 1MJC | 69 | 28 (3, 25) | 3-182 | 4.36e+26 | 1.2e+03 |
| 1CSP | 67 | 30 (6, 24) | 3-182 | 5.02e+30 | 5.8e+02 |
| 1BK2 | 57 | 24 (11, 13) | 3-182 | 1.18e+32 | 8.2e+02 |
| 1SHG | 57 | 28 (9, 19) | 3-182 | 2.13e+32 | 4.6e+02 |
| 1CSK | 58 | 30 (9, 21) | 3-49 | 4.09e+32 | 7.0e+02 |
| 1SHF | 59 | 30 (9, 21) | 3-56 | 1.05e+34 | 5.5e+02 |
| 1FYN | 62 | 23 (13, 10) | 3-186 | 5.04e+36 | 5.1e+03 |
| 1PIN | 154 | 28 (11, 17) | 3-194 | 5.32e+39 | 1.0e+05 |
| 1NXB | 62 | 34 (9, 25) | 3-56 | 2.61e+41 | 5.5e+02 |
| 1TEN | 90 | 39 (11, 28) | 3-66 | 6.17e+43 | 5.8e+03 |
| 1POH | 85 | 46 (3, 43) | 3-182 | 8.02e+43 | 5.5e+03 |
| 2DRI | 271 | 37(10,27) | 3-186 | 1.16e+47 | 8.5e+04 |
| 1FNA | 91 | 38 (15, 23) | 3-48 | 3.02e+47 | 3.7e+03 |
| 1UBI | 76 | 40 (14, 26) | 3-182 | 2.43e+49 | 1.6e+03 |
| 1C9O | 66 | 43 (11, 32) | 3-182 | 3.77e+49 | 7.3e+03 |
| 1CTF | 68 | 39 (20, 19) | 3-56 | 3.95e+51 | 8.7e+02 |
| 2PCY | 99 | 46 (16, 30) | 3-56 | 2.34e+52 | 5.5e+03 |
| 1DKT | 71 | 46 (14, 32) | 3-190 | 3.94e+58 | 1.2e+03 |
| 2TRX | 108 | 61 (6, 55) | 3-186 | 9.02e+59 | 1.1e+05 |
| 1CM1 | 143 | 42(18,24) | 3-186 | 3.73e+63 | 2.3e+04 |
| 1BRS | 108 | 44 (19, 25) | 3-194 | 1.67e+64 | 5.5e+05 |
| 1CDL | 142 | 40 (21, 19) | 3-186 | 5.68e+65 | 2.0e+05 |
| 1LZ1 | 130 | 59 (22, 37) | 3-57 | 1.04e+72 | 5.0e+04 |
| 1GVP | 87 | 52 (28, 24) | 3-182 | 1.51e+78 | 1.6e+05 |
| 1RIS | 97 | 56 (27, 29) | 3-182 | 1.23e+80 | 7.6e+03 |
| 2RN2 | 155 | 69 (27, 42) | 3-66 | 3.68e+80 | 2.3e+05 |
| 1CSE | 274 | 97 (4, 93) | 3-183 | 8.35e+82 | 2.9e+06 |
| 1HNG | 175 | 85 (13, 72) | 3-182 | 3.70e+88 | 5.4e+05 |
| 3CHY | 128 | 74 (31, 43) | 3-66 | 2.36e+92 | 3.8e+04 |
| 1L63 | 162 | 83 (21, 62) | 3-182 | 2.17e+94 | 3.2e+05 |

[1] Traoré S, Allouche D, André I, de Givry S, Katsirelos G, Schiex T, Barbe S (2013) A new framework for computational protein design through cost function network optimization. Bioinformatics. 29(17):2129-2136

|  |
| --- |
| **Table S-2** Comparison of search strategies for solving the GMEC identification problem. For each method are reported the number of nodes expanded and the number of backtracks (nd (bt)), the CPU-time in second (time) and the speed (nd/min). Both DFBB and A\*-CFN use default ordering heuristics and initial LDS upper bounding.  |
| PDB | Sequence Conformation Space Size | *A\**-vanilla | *A\**-CFN | DFBB |
| nd | time (s) | nd/min | nd | time (s) | nd/min | nd(bt) | time (s) | nd/min |
| 1MJC | 4.36E+26 | 0 | 5.1 | - | 0 | 2.8 | - | 0 | 6.0 | - |
| 1CSP | 5.02E+30 | 127 | 94.6 | 80 | 4 | 40.4 | 5 | 82(56) | 97.0 | 50 |
| 1BK2 | 1.18E+32 | 5,234 | 285.3 | 1,100 | 11 | 20.0 | 33 | 144(90) | 9.1 | 949 |
| 1SHG | 2.13E+32 | 558 | 77.5 | 432 | 24 | 76.6 | 18 | 114(75) | 29.5 | 231 |
| 1CSK | 4.09E+32 | 113 | 7.8 | 869 | 13 | 8.1 | 96 | 38(15) | 3.6 | 633 |
| 1SHF | 1.05E+34 | 21 | 13.9 | 90 | 6 | 14.0 | 25 | 32(13) | 14.0 | 137 |
| 1FYN | 5.04E+36 | 58,380 | 3,355.0 | 1,044 | 47 | 253.0 | 11 | 575(425) | 159.4 | 216 |
| 1PIN | 5.32E+39 | 24 | 1,480.4 | 0 | 8 | 580.8 | 0 | 34(13) | 1,465.5 | 1 |
| 1NXB | 2.61E+41 | 45 | 3.8 | 710 | 0 | 1.9 | - | 18(0) | 3.6 | 300 |
| 1TEN | 6.17E+43 | 243 | 37.1 | 392 | 5 | 37.8 | 7 | 45(18) | 37.8 | 71 |
| 1POH | 8.02E+43 | 169 | 29.8 | 340 | 9 | 29.2 | 18 | 82(32) | 11.4 | 431 |
| 2DRI | 1.16E+47 | - | - | - | - | - | - | - | - | - |
| 1FNA | 3.02E+47 | 109,523 | 638.6 | 10,290 | 344 | 296.0 | 69 | 402(287) | 61.7 | 390 |
| 1UBI | 2.43E+49 | 193,077 | 6,895.0 | 1,680 | 401 | 1,447.2 | 16 | 764(659) | 1,297.9 | 35 |
| 1C9O | 3.77E+49 | 63,974 | 960.6 | 3,995 | 44 | 246.6 | 10 | 275(221) | 242.1 | 68 |
| 1CTF | 3.95E+51 | 22,658 | 2,363.0 | 575 | 1027 | 2,254.9 | 27 | 708(567) | 1,815.1 | 23 |
| 2PCY | 2.34E+52 | 14,929 | 299.1 | 2,994 | 21 | 28.1 | 44 | 114(67) | 69.6 | 98 |
| 1DKT | 3.94E+58 | 202,611 | 1,078.6 | 11,270 | 23 | 959.8 | 1 | 174(74) | 324.7 | 32 |
| 2TRX | 9.02E+59 | 29,845 | 123.5 | 14,499 | 44 | 131.7 | 20 | 178(83) | 128.9 | 82 |
| 1CM1 | 3.73E+63 | - | - | - | - | - | - | - | - | - |
| 1BRS | 1.67E+64 | - | - | - | - | - | - | - | - | - |
| 1CDL | 5.68E+65 | - | - | - | - | - | - | - | - | - |
| 1LZ1 | 1.04E+72 | - | - | - | 531 | 918.3 | 34 | 569(376) | 494.6 | 69 |
| 1GVP | 1.51E+78 | - | - | - | 280 | 8,099.8 | 2 | - | - | - |
| 1RIS | 1.23E+80 | - | - | - | - | - | - | - | - | - |
| 2RN2 | 3.68E+80 | 15,845 | 1,276.7 | 744 | 51 | 3,845.2 | 0 | 364(258) | 1,139.9 | 19 |
| 1CSE | 8.35E+82 | 26 | 79.0 | 19 | 0 | 187.2 | - | 22(0) | 185.9 | 7 |
| 1HNG | 3.70E+88 | 16 | 1,739.0 | 0 | 0 | 1,721.7 | - | 37(14) | 1,724.1 | 1 |
| 3CHY | 2.36E+92 | - | - | - | - | - | - | 18,739(16,110) | 3,685.2 | 305 |
| 1L63 | 2.17E+94 | 19,567 | 1,409.9 | 832 | 24 | 1,351.9 | 1 | 231(138) | 1,460.8 | 9 |

|  |
| --- |
| **Table S-3** Assessment of branching schemes for solving the GMEC identification problem. For each branching scheme are reported the number of nodes expanded and the number of backtracks (nd (bt)) and the CPU-time in second (time). Dashed lines correspond to cases where no solution could be found. |
| PDB | Sequence Conformation Space Size | Binary | N-ary | Dichotomic | SCP |
| nd(bt) | time (s) | nd(bt) | time (s) | nd(bt) | time (s) | nd(bt) | time (s) |
| 1MJC | 4.36E+26 | 21(0) | 0.7 | 21(0) | 0.8 | 21(0) | 0.8 | 21(0) | 1.2 |
| 1CSP | 5.02E+30 | 215(96) | 36.1 | 5,216(5,105) | 81.8 | 269(123) | 26.2 | 67(22) | 14.3 |
| 1BK2 | 1.18E+32 | 92(35) | 8.9 | 234(183) | 8.6 | 62(20) | 8.1 | 68(23) | 8.6 |
| 1SHG | 2.13E+32 | 199(87) | 7.0 | 344(245) | 11.7 | 171(73) | 6.1 | 123(49) | 3.4 |
| 1CSK | 4.09E+32 | 27(1) | 1.6 | 27(1) | 1.5 | 27(1) | 2.3 | 27(1) | 2.5 |
| 1SHF | 1.05E+34 | 45(9) | 1.5 | 50(14) | 1.5 | 29(1) | 2.2 | 29(1) | 2.2 |
| 1FYN | 5.04E+36 | 187(83) | 212.4 | 1,916(1,839) | 384.2 | 209(94) | 303.5 | 75(27) | 88.6 |
| 1PIN | 5.32E+39 | 376(175) | 438.8 | 5,908(5,719) | 318.6 | 0(0) | - | 250(112) | 141.2 |
| 1NXB | 2.61E+41 | 35(1) | 1.4 | 35(1) | 1.4 | 35(1) | 1.4 | 35(1) | 1.4 |
| 1TEN | 6.17E+43 | 117(41) | 4.1 | 151(81) | 2.6 | 77(21) | 8.6 | 73(19) | 2.2 |
| 1POH | 8.02E+43 | 69(14) | 4.9 | 94(39) | 4.7 | 61(10) | 4.6 | 57(8) | 4.8 |
| 2DRI | 1.16E+47 | 4,734(2,350) | 6,825.2 | 0(0) | - | 0(0) | - | 4,448(2,207) | 4,426.2 |
| 1FNA | 3.02E+47 | 131(48) | 27.6 | 209(140) | 23.9 | 109(37) | 15.2 | 165(65) | 19.6 |
| 1UBI | 2.43E+49 | 1,183(572) | 309.1 | 2,901(2,620) | 475.5 | 1,041(501) | 198.4 | 807(384) | 192.4 |
| 1C9O | 3.77E+49 | 201(82) | 131.1 | 1,627(1,518) | 240.1 | 141(52) | 77.5 | 95(29) | 50.2 |
| 1CTF | 3.95E+51 | 264(114) | 103.3 | 782(665) | 137.8 | 226(95) | 96.4 | 332(148) | 116.2 |
| 2PCY | 2.34E+52 | 115(37) | 9.9 | 291(213) | 6.1 | 91(25) | 8.1 | 107(33) | 5.8 |
| 1DKT | 3.94E+58 | 236(95) | 172.4 | 1,192(1,075) | 191.9 | 186(70) | 297.1 | 134(44) | 109.5 |
| 2TRX | 9.02E+59 | 204(72) | 58.0 | 567(448) | 34.2 | 190(65) | 47.5 | 174(57) | 24.8 |
| 1CM1 | 3.73E+63 | 537(248) | 6,097.2 | 1,657(1,407) | 5,412.0 | 501(230) | 6,070.4 | 507(233) | 5,218.5 |
| 1BRS | 1.67E+64 | 0(0) | - | 0(0) | - | 0(0) | - | 0(0) | - |
| 1CDL | 5.68E+65 | 0(0) | - | 0(0) | - | 0(0) | - | 0(0) | - |
| 1LZ1 | 1.04E+72 | 1,367(657) | 210.9 | 3,305(2,656) | 413.0 | 1,237(592) | 116.9 | 415(181) | 51.3 |
| 1GVP | 1.51E+78 | 0(0) | - | 0(0) | - | 0(0) | - | 0(0) | - |
| 1RIS | 1.23E+80 | 0(0) | - | 0(0) | - | 0(0) | - | 0(0) | - |
| 2RN2 | 3.68E+80 | 372(155) | 54.5 | 711(530) | 46.8 | 290(114) | 38.4 | 280(109) | 26.0 |
| 1CSE | 8.35E+82 | 206(62) | 10.3 | 480(337) | 17.7 | 100(9) | 9.1 | 100(9) | 4.7 |
| 1HNG | 3.70E+88 | 612(266) | 287.3 | 3,476(3,115) | 398.3 | 454(187) | 154.5 | 164(42) | 78.5 |
| 3CHY | 2.36E+92 | 0(0) | - | 0(0) | - | 0(0) | - | 11,565(5,746) | 4,974.4 |
| 1L63 | 2.17E+94 | 229(75) | 187.3 | 914(766) | 179.9 | 221(71) | 124.7 | 263(92) | 98.7 |

|  |
| --- |
| **Table S-4** The sequence conformation enumeration problem. For each method are reported the number of nodes expanded (nd) and the CPU-time in second (time). Dashed lines correspond to cases where no solution could be found. |
| PDB | Sequence Conformation Space Size | *A\**-CFN | DFBB | *A\**-SCP | DFBB-SCP |
| nd | time (s) | nd | time (s) | nd | time (s) | nd | time (s) |
| 1MJC | 4.36E+26 | 4,408,541 | 232.6 | 4,537,935 | 140.7 | 4,415,342 | 143.7 | 4,555,451 | 178.7 |
| 1CSP | 5.02E+30 | 244,640 | 1,474.7 | 252,247 | 883.1 | 252,778 | 397.4 | 269,065 | 115.3 |
| 1BK2 | 1.18E+32 | 657,753 | 153.3 | 666,956 | 68.0 | 690,3 | 425.8 | 741,604 | 677.1 |
| 1SHG | 2.13E+32 | 1,326,902 | 418.8 | 1,410,363 | 383.6 | 1,330,026 | 78.8 | 1,400,933 | 85.9 |
| 1CSK | 4.09E+32 | 922,038 | 262.3 | 948,083 | 308.1 | 928,305 | 44.0 | 944,657 | 32.2 |
| 1SHF | 1.05E+34 | 66,848 | 191.3 | 71,275 | 6.3 | 64,703 | 29.0 | 68,171 | 42.4 |
| 1FYN | 5.04E+36 | - |  | 842,041 | 661.9 | 941,879 | 2,418.9 | 1,042,143 | 2,406.8 |
| 1PIN | 5.32E+39 | - |  | - |  | 36,384 | 7,431.1 | 39,228 | 2,354.6 |
| 1NXB | 2.61E+41 | - |  | 97,333,939 | 5,949.1 | - |  | 97,055,693 | 5,916.9 |
| 1TEN | 6.17E+43 | 14,222,924 | 2,191.7 | 14,407,481 | 950.2 | 13,902,152 | 1,665.0 | 14,558,687 | 2,129.0 |
| 1POH | 8.02E+43 | - |  | - |  | - |  | - |  |
| 2DRI | 1.16E+47 | - |  | - |  | - |  | - |  |
| 1FNA | 3.02E+47 | - |  | 208,370,115 | 5,621.2 | - |  | - |  |
| 1UBI | 2.43E+49 | - |  | 744,735 | 8,283.8 | - |  | - |  |
| 1C9O | 3.77E+49 | - |  | 8,030,665 | 2,157.5 | 7,293,462 | 2,448.3 | 8,153,909 | 574.0 |
| 1CTF | 3.95E+51 | - |  | 40,391,526 | 4,202.1 | 40,064,876 | 3,428.2 | 41,465,042 | 2,501.9 |
| 2PCY | 2.34E+52 | 479,406 | 181.5 | 481,801 | 189.9 | 471,664 | 106.7 | 478,295 | 33.5 |
| 1DKT | 3.94E+58 | - |  | - |  | - |  | - |  |
| 2TRX | 9.02E+59 | - |  | - |  | - |  | - |  |
| 1CM1 | 3.73E+63 | - |  | - |  | - |  | - |  |
| 1BRS | 1.67E+64 | - |  | - |  | - |  | - |  |
| 1CDL | 5.68E+65 | - |  | - |  | - |  | - |  |
| 1LZ1 | 1.04E+72 | - |  | 21,137,757 | 4,705.1 | 20,240,02 | 2,809.5 | 21,310,987 | 4,247.3 |
| 1GVP | 1.51E+78 | - |  | - |  | - |  | - |  |
| 1RIS | 1.23E+80 | - |  | - |  | - |  | - |  |
| 2RN2 | 3.68E+80 | - |  | - |  | 56,012,326 | 7,570.1 | 60,037,744 | 2,790.0 |
| 1CSE | 8.35E+82 | - |  | - |  | - |  | - |  |
| 1HNG | 3.70E+88 | - |  | - |  | - |  | - |  |
| 3CHY | 2.36E+92 | - |  | - |  | - |  | - |  |
| 1L63 | 2.17E+94 | - |  | - |  | - |  | - |  |

|  |
| --- |
| **Table S-5** Enumeration of unique sequences. For each method are reported the number of nodes expanded (nd) and the CPU-time in second (time). Dashed lines correspond to cases where no solution could be found. |
| PDB | Sequence Conformation Space Size | *A\**-SCP-unique | DFBB-SCP-unique |
| nd | time (s) | nd | time (s) |
| 1MJC | 4.36E+26 | 2,187 | 37.4 | 181 | 5.6 |
| 1CSP | 5.02E+30 | 9,875 | 152.7 | 1,734 | 74.5 |
| 1BK2 | 1.18E+32 | 60,462 | 890.8 | 23,801 | 436.6 |
| 1SHG | 2.13E+32 | 10,567 | 26.1 | 1,357 | 28.0 |
| 1CSK | 4.09E+32 | 3,645 | 5.5 | 415 | 5.0 |
| 1SHF | 1.05E+34 | 605 | 179.5 | 85 | 43.3 |
| 1FYN | 5.04E+36 | 85,145 | 846.5 | 14,530 | 2,241.2 |
| 1PIN | 5.32E+39 | 4,328 | 2,283.4 | 2,199 | 6,229.5 |
| 1NXB | 2.61E+41 | 13,477 | 459.1 | 1,088 | 24.1 |
| 1TEN | 6.17E+43 | 7,337 | 136.2 | 751 | 113.5 |
| 1POH | 8.02E+43 | 7,129 | 137.7 | 364 | 93.9 |
| 2DRI | 1.16E+47 | - |  | - |  |
| 1FNA | 3.02E+47 | 97,585 | 807.6 | 8,261 | 908.0 |
| 1UBI | 2.43E+49 | 26,179 | 7,767.5 | 20,833 | 8,627.1 |
| 1C9O | 3.77E+49 | 34,781 | 2,100.6 | 3,419 | 1,196.7 |
| 1CTF | 3.95E+51 | 463,902 | 2,243.2 | 67,465 | 3,270.8 |
| 2PCY | 2.34E+52 | 2,677 | 28.2 | 339 | 42.9 |
| 1DKT | 3.94E+58 | - |  | 75,586 | 5,080.7 |
| 2TRX | 9.02E+59 | 8,536 | 101.1 | 728 | 270.0 |
| 1CM1 | 3.73E+63 | - |  | - |  |
| 1BRS | 1.67E+64 | - |  | - |  |
| 1CDL | 5.68E+65 | - |  | - |  |
| 1LZ1 | 1.04E+72 | 21,311 | 1,574.1 | 5,450 | 3,653.2 |
| 1GVP | 1.51E+78 | - |  | - |  |
| 1RIS | 1.23E+80 | - |  | - |  |
| 2RN2 | 3.68E+80 | 108,496 | 1,075.5 | 13,348 | 1,691.8 |
| 1CSE | 8.35E+82 | 3,356 | 67.7 | 1,707 | 1,429.3 |
| 1HNG | 3.70E+88 | 47,438 | 3,006.0 | - |  |
| 3CHY | 2.36E+92 | - |  | - |  |
| 1L63 | 2.17E+94 | - |  | - |  |

|  |
| --- |
| **Table S-6** Cross comparison of search methods for the generation of sub-optimal sequence-conformation ensembles as well as enumeration of sub-optimal unique sequences. |
|  | *A\**-CFN | DFBB | *A\**-SCP | DFBB-SCP | *A\**-SCP-unique |
| DFBB | -2239.7 |  |   |  |  |
| 13/15 |  |  |  |  |
| *A\**-SCP | -2150.7 | 205.1 |  |  |  |
| 13/14 | 8/17 |  |  |  |
| DFBB-SCP | -2934.7 | -613.2 | -927.4 |  |  |
| 14/15 | 10/17 | 9/15 |  |  |
| *A\**-SCP-unique | -4675.2 | -3075.4 | -3241.4 | -2579.0 |  |
| 20/21 | 18/21 | 19/21 | 17/21 |  |
| DFBB-SCP-unique | -4125.7 | -2525.9 | -2691.9 | -2131.0 | 549.5 |
| 20/21 | 17/21 | 18/21 | 16/20 | 9/21 |
| In each cell, the upper number is an averaged runtime difference between the row and the column strategy. Hence, a negative value is in favor of the search method in the cell’s row. In the lower part of the cell is indicated the number of instances solved faster by the method of the cell’s row (the denominator is the total number of instances considered).The runtime assigned to unsolved instances is 9,000 sec (which is a runtime lower bound for unsolved cases). Only cases for which the runtime difference is greater than 1 sec. are considered. |

*A\**-CFN is outperformed by DFBB in 13 cases out of 15 (only 8 were actually solved by *A\**-CFN while DFBB solved all 15). *A\**-SCP solved a total of 14 cases and outperformed *A\**-CFN in 13 out of 14 cases. However, DFBB turned out to be slightly more efficient (an average runtime gain of 205.1 sec in favor of DFBB and 9 out of 17 cases solved faster). Ultimately, DFBB-SCP outperforms *A\**-CFN, DFBB, and *A\**-SCP. Hence, SCP improves both *A\**-CFN and DFBB search strategies. A significant gain in runtime is always obtained when unique sequence SCP is activated (more than 2,000 sec in average and up to 4675.2 sec). [↑](#endnote-ref-1)