**Supplementary Material**

**3D representations of amino acids – applications to protein sequence comparison and classification**

Jie Lia and Patrice Koehlb

a Genome Center, University of California, Davis, 451 Health Sciences Drive, Davis, CA 95616. E-mail: [jjsli@ucdavis.edu](mailto:jjsli@ucdavis.edu)

b Department of Computer Science and Genome Center, University of California, Davis, One Shields Ave, Davis, CA, 95616. E-mail: [koehl@cs.ucdavis.edu](mailto:koehl@cs.ucdavis.edu)

**Table S1a:** 62 proteins in CATH class 1.10.10 included in CATH605

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 1i27A0 | 1hw5B2 | 2irfL0 | 1mnmD0 | 1b4aF1 | 1d5yD2 | 2bby00 | 1dpuA0 |
| 1dp7P0 | 1ois01 | 1qbjC0 | 1akhA0 | 1tc3C0 | 1octC2 | 1wjfB0 | 1d8kA0 |
| 1b6a02 | 1aa7B2 | 1hcrA0 | 1gln05 | 1cf7B0 | 1hstB0 | 1tnt00 | 1d3yB1 |
| 1mgtA2 | 1bm9B0 | 1bjaB0 | 1bl0A2 | 1cf7A0 | 1gdtB3 | 1msfC2 | 1bw6A0 |
| 1dtr01 | 1fnnB3 | 2occU0 | 1bl0A1 | 1mmsB0 | 1uaaB2 | 1msfC1 | 1ba500 |
| 1fjlC0 | 1opc00 | 1bia01 | 1ft9B2 | 1repC2 | 2hfh00 | 1leb00 | 1aoy00 |
| 1b9mB1 | 1smtB0 | 1f1zB2 | 1pdnC2 | 1i50J0 | 2ezl00 | 1fc3C0 |  |
| 2tct01 | 1duxF0 | 1b72B0 | 1pdnC1 | 1lfb00 | 2ezi00 | 1erd00 |  |

**Table S1b:** 169 proteins from CATH 2.60.40 in CATH605

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 1mfmA0 | 1nif02 | 1bec02 | 1soxB3 | 1avaB2 | 1cdy01 | 1bvzB3 | 1a02N2 |
| 1f86B0 | 1nif01 | 1bec01 | 1vcaB2 | 1edhB2 | 1qfhB2 | 1bvzB1 | 1a02N1 |
| 1aac00 | 1c5cL2 | 1nbcB0 | 1vcaB1 | 1dqtD0 | 1iam02 | 1iraY3 | 1smaB3 |
| 1plc00 | 1c5cH2 | 1wapV0 | 1tf4B2 | 1uok03 | 1iam01 | 1iraY2 | 1dr9A2 |
| 1eajB0 | 1gof03 | 1nkr02 | 7taa02 | 1amx00 | 1eerC2 | 1iraY1 | 1dr9A1 |
| 1i4fB0 | 1fzkA2 | 1yaiC0 | 1bf203 | 1bquB2 | 1eerC1 | 1rlw00 | 1kcw06 |
| 1bqk00 | 1dp0D4 | 1svb04 | 1bf201 | 1bquB1 | 1acc04 | 4kbpD1 | 1kcw05 |
| 1qtsA1 | 1dp0D2 | 1rsy00 | 1h5bD0 | 1fyhE1 | 1cyx00 | 1djsA1 | 1bhgB2 |
| 2mcm00 | 1qba04 | 1tvdB0 | 1cfb02 | 2fbjH2 | 1qpxB2 | 1ejfB0 | 1shsH0 |
| 1dlfL0 | 1qba01 | 1gcyA2 | 1cfb01 | 1f13B4 | 1ycsA0 | 1tcrA2 | 1qunP2 |
| 1dlfH0 | 1ibyD0 | 1epfD2 | 1ayoB0 | 1f13B3 | 1iarB1 | 1tcrA1 | 1cid02 |
| 1e30B0 | 2cbp00 | 1epfD1 | 1f00I2 | 1axiB2 | 1cwvA4 | 1qmuA2 | 1cid01 |
| 1jzgA0 | 1e42B1 | 1cvrA2 | 1f00I1 | 1bftB0 | 1cwvA3 | 1pdkB0 | 2xbd00 |
| 1edqA1 | 1aohB0 | 1wwcA0 | 1aozB3 | 1fnf02 | 1cwvA2 | 1bqsA2 | 2fnbA0 |
| 1f0lB3 | 2fcbA2 | 1cs6A4 | 1aozB2 | 1dceC2 | 1i1rA1 | 1bqsA1 | 1wiu00 |
| 1ej8A0 | 2fcbA1 | 1cs6A3 | 1aozB1 | 1f97A2 | 1bag02 | 1cd1C2 | 1tiu00 |
| 1qhoA4 | 1fna00 | 1cs6A2 | 1clc01 | 1f97A1 | 1bg1A2 | 1fo0B0 | 1ksr00 |
| 1qhoA3 | 1fltY0 | 1cs6A1 | 1fngC2 | 1qg3B2 | 1cd800 | 1mspB0 | 1g84A0 |
| 1qhoA2 | 1dqiD0 | 1who00 | 1dfx00 | 1qg3B1 | 1xbrB0 | 1vcbL2 | 1exh00 |
| 1smd02 | 1bli03 | 1cczA2 | 1hoe00 | 1bdyB0 | 1dn2B1 | 1cf1D1 | 1bpv00 |
| 1e2wB1 | 1ten00 | 1cczA1 | 1cdcB0 | 1cdy02 | 1eut02 | 1egjA0 | 1b4rA0 |
| 1ahm00 |  |  |  |  |  |  |  |

**Table S1c:** 67 proteins from CATH 3.20.20 in CATH605

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 7a3hA0 | 1h61A0 | 1frb00 | 1dxeB0 | 1bd0B1 | 1bf202 | 1exbA0 | 1bag01 |
| 1qtwA0 | 1thfD0 | 1jae01 | 1mucB2 | 1pud00 | 1pii02 | 1rpxC0 | 1bvzB2 |
| 1fxmA0 | 1ccwD1 | 1burD2 | 1oneB2 | 1f8mD0 | 1euaC0 | 1eceB0 | 1hg3H0 |
| 1e4mM0 | 1f74C0 | 2bvwB0 | 1nar00 | 1onrB0 | 1ec7D2 | 1dqwD0 | 1ex1A1 |
| 2tpsB0 | 1lucB0 | 1ezwA0 | 1epxD0 | 1fxqB0 | 1a4mD0 | 1d7kB2 |  |
| 1qnrA0 | 1edg00 | 1amk00 | 1uroA0 | 1eswA0 | 2plc00 | 1dik04 |  |
| 1i0dB0 | 1dosB0 | 1dp0D3 | 1pymB0 | 1a5300 | 1qfeB0 | 1dbtC0 |  |
| 4xis00 | 1cnv00 | 1qba03 | 1eokA0 | 1aq0B0 | 1ho1D0 | 1qpoF2 |  |
| 1dvjD0 | 1qhoA1 | 1eqcA0 | 1byb00 | 2ebn00 | 7reqD1 | 1qvbB0 |  |

**Table S1d**: 92 proteins from CATH 3.30.70 in CATH605

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 1cbn00 | 1hfeM2 | 2acy00 | 1urnC0 | 1fd200 | 1ek8A2 | 1fsz02 | 1u2fA0 |
| 2fdn00 | 1cs1D2 | 1fehA3 | 1xer00 | 1bjnB2 | 1g51B3 | 1e7kB0 | 1qm9A2 |
| 1cc8A0 | 1ctf00 | 1ftrD2 | 1bxyB0 | 1b9hA2 | 1b4bC0 | 1fjgP0 | 1qm9A1 |
| 1eqoA0 | 1burD1 | 1lbu02 | 1scjB0 | 1fxrB0 | 1dcpH0 | 1fjgJ0 | 1iba00 |
| 1qu9C0 | 1vjw00 | 1cqmB0 | 1alo03 | 1dfoD2 | 1cliD1 | 2u2fA0 | 1gh8A0 |
| 1mroE1 | 1qd1B2 | 1hlwA0 | 1lvk06 | 1brwB1 | 1psdB3 | 2u1a00 | 1fjeB2 |
| 1mroD2 | 1f60B0 | 1bhp00 | 1f0xB4 | 1nsa01 | 1bjt04 | 2tbd00 | 1fjeB1 |
| 1dbfC0 | 2pii00 | 1fxlA2 | 1el6C3 | 1d09D1 | 1dar05 | 2hqi00 | 1cn9A0 |
| 1mla01 | 1aye01 | 1qupB1 | 1ffgD0 | 1b3tB0 | 1pysB6 | 2fxb00 |  |
| 1dj0B2 | 2bopA0 | 1ha102 | 1f08B0 | 1xxaF0 | 1qaxB1 | 2fmr00 |  |
| 1dj0B1 | 1vhh00 | 1dqaD3 | 1seiB1 | 1cg2D2 | 1c4kA3 | 2aw000 |  |
| 1ptf00 | 1kp6A0 | 1regY0 | 1f3vA0 | 1b0pB5 | 1cvjH1 | 1vih00 |  |

**Table S1e**: 215 proteins from CATH 3.40.50 in CATH605

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 1gci00 | 1b0uA0 | 4uagA1 | 1chd00 | 1bgvA1 | 1g3qA0 | 1a7701 | 2liv02 |
| 1g66A0 | 1qczA0 | 4lipE0 | 1cqxB3 | 1b93C0 | 1nbaD0 | 1pvdB3 | 1hwxF3 |
| 1byi00 | 8abp02 | 2nacB2 | 2nsyB0 | 1trkB3 | 1qorB2 | 1qo0E1 | 1nksF0 |
| 1cex00 | 1fj2B0 | 1f60A1 | 1cf9D3 | 1trkB2 | 2pgd01 | 1f8yB0 | 1ndh02 |
| 1qj4A0 | 1a4iB2 | 1pdo00 | 1dusA0 | 1trkB1 | 1efdN1 | 1c7qA2 | 1ei9A0 |
| 1heuB2 | 1bfd03 | 1din00 | 1b8pA1 | 1qnf01 | 1efvB0 | 2hgsA4 | 1an9B1 |
| 1qlwB0 | 1bfd02 | 1fdr02 | 1e6cB0 | 1ej2A0 | 1efvA2 | 3pmgB2 | 1e79F2 |
| 1b6g00 | 1bfd01 | 1gsoA1 | 1dqpB0 | 1dv1B1 | 1efvA1 | 1a5t01 | 1b74A2 |
| 1hdoA0 | 1ej0A0 | 1fbnA0 | 4pgaB2 | 1bif02 | 1bd3D0 | 1d5wC0 | 1qhxA0 |
| 1e58A0 | 1h2rS1 | 1atzB0 | 4pgaB1 | 1bif01 | 1ecpF0 | 1a1vA1 | 1qhhA0 |
| 1d4oA0 | 1iatA2 | 1qjcB0 | 1d9tF0 | 1amuB2 | 1cp2B0 | 1fuiF1 | 1epuA1 |
| 1qtnA0 | 1e2uA4 | 1duvI2 | 1d4aD0 | 1djqB2 | 1esc00 | 2rslC0 | 1egaB1 |
| 1f4pA0 | 1e2uA3 | 1dozA2 | 1eyzB1 | 1whtA0 | 1lci02 | 1cjsA2 | 1kmmD2 |
| 2pth00 | 1b8oA0 | 1iibB0 | 1jkmB0 | 1i7qD0 | 1gpmD1 | 1ehiB1 | 1adjD2 |
| 1ctqA0 | 1deoA0 | 1yacB0 | 1f0yB1 | 1ffh02 | 1h9aA1 | 1qtrA0 | 1b3rD2 |
| 1oaa00 | 1zin00 | 1jfrB0 | 1fukA0 | 1cozB0 | 1af702 | 1bykB2 | 1bmtB2 |
| 1ga4A0 | 1e9eA0 | 3tgl00 | 1ejbE0 | 1vid00 | 7reqD2 | 1b0pB6 | 2hpaD0 |
| 1c1dB2 | 1moq02 | 1toaB2 | 1bu8A1 | 1f0kB2 | 1fnnB2 | 1b0pB2 | 1atiB2 |
| 1es9A0 | 1qrrA0 | 1toaB1 | 1iow01 | 4tmkA0 | 1fmtB1 | 1b0pB1 | 1fsz01 |
| 1brt00 | 1qb7A0 | 1d2nA1 | 1bg601 | 1g38D1 | 3minD3 | 1f48A2 | 1tdj02 |
| 1fnr02 | 16pk02 | 1dqsB1 | 1c3jA2 | 1fcjD2 | 3minD2 | 1do8D2 | 1d6mA1 |
| 1qsgH0 | 16pk01 | 1ido00 | 1ekxC2 | 1eudB3 | 3minD1 | 1cfzF0 | 1g2iC0 |
| 1tc1B0 | 1qmgD1 | 1zpdF2 | 1a9xH2 | 1eudA2 | 3minC3 | 1qfjD2 | 1ex1A2 |
| 1dqzB0 | 1evlD2 | 1zpdF1 | 1a9xG8 | 1eudA1 | 3minC2 | 1c4xA0 | 1eiwA0 |
| 1ccwC0 | 1c4oA3 | 1ldg01 | 1a9xG5 | 1ad202 | 1dhr00 | 1qs0B2 | 1e5dB1 |
| 1tca00 | 1gegH0 | 1dljA2 | 1poxB2 | 1sur00 | 1e20A0 | 1qs0B1 | 1b6sD1 |
| 1i6wB0 | 1d1qB0 | 1dljA1 | 1poxB1 | 2pia02 | 1pjcA2 | 1qs0A0 |  |

**Table S2:** Coordinates of the N-dimensional vectors representing the amino acids, based on BLOSUM62

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Coordinates a) | | | | |
| Amino acid | 1 | 2 | 3 | 4 | 5 |
| Ala (A) | 0.189 | -3.989 | 1.989 | 0.140 | 1.009 |
| Arg (R) | 5.007 | 0.834 | -2.709 | -2.027 | 3.696 |
| Asn (N) | 7.616 | 0.943 | 0.101 | 3.308 | 0.207 |
| Asp (D) | 7.781 | 0.030 | 1.821 | 1.376 | -3.442 |
| Cys (C) | -5.929 | -4.837 | 6.206 | 2.884 | 5.365 |
| Gln (Q) | 5.480 | 1.293 | -3.091 | -2.348 | 1.628 |
| Glu (E) | 7.444 | 1.005 | -2.121 | -1.307 | -1.011 |
| Gly (G) | 4.096 | 0.772 | 7.120 | 0.211 | -1.744 |
| His (H) | 3.488 | 6.754 | -2.703 | 4.989 | 0.452 |
| Ile (I) | -7.883 | -4.900 | -2.230 | 0.990 | -2.316 |
| Leu (L) | -7.582 | -3.724 | -2.740 | -0.736 | -0.208 |
| Lys (K) | 5.665 | -0.166 | -2.643 | -2.808 | 2.474 |
| Met (M) | -5.200 | -2.547 | -3.561 | -1.730 | 0.859 |
| Phe (F) | -8.681 | 4.397 | -0.732 | 1.883 | -1.987 |
| Pro (P) | 4.281 | -2.932 | 2.319 | -3.269 | -4.451 |
| Ser (S) | 4.201 | -1.948 | 1.453 | 1.226 | 1.014 |
| Thr (T) | 0.774 | -3.192 | 0.666 | 0.070 | 0.407 |
| Trp (W) | -8.492 | 9.958 | 4.874 | -5.288 | 0.672 |
| Tyr (Y) | -6.147 | 7.590 | -2.065 | 2.413 | -0.562 |
| Val (V) | -6.108 | -5.341 | -1.953 | 0.025 | -2.062 |

a) Coordinates along the principal components of the BLOSUM62 matrix. The first three components defined the 3D vectors representing the amino acids.