# Supplementary Information: Amino acid composition of proteins reduces deleterious impact of mutations

Sahand Hormoz[1, 2]

[1] *School of Engineering and Applied Sciences and*
*Kavli Institute for Bionano Science and Technology,*
*Harvard University, Cambridge, Massachusetts 02138, USA*

[2] *Department of Cancer Biology, Dana–Farber*
*Cancer Institute, Boston, Massachusetts 02115*

## SUPPLEMENTARY TABLE 1

**Proteome List:** The following 75 organisms where used as the set of natural proteomes. Their complete non-redundant proteome sets were downloaded from UniProt Database (http://www.uniprot.org/) UniParc Archives [1]. The optimal growth temperatures (OGT) in units of °C were obtained from [2].

| **Organism** | **OGT** |
|---|---|
| Acidobacteria bacterium Ellin345 | 25 |
| Aeropyrum pernix | 95 |
| Anabaena variabilis ATCC 29413 | 35 |
| Aquifex aeolicus | 85 |
| Agrobacterium tumefaciens C58 UWash | 26 |
| Archaeoglobus fulgidus | 83 |
| Bacillus anthracis Ames | 30 |
| Bacillus licheniformis DSM 13 | 37 |
| Bordetella bronchiseptica | 36 |
| Bdellovibrio bacteriovorus | 30 |
| Campylobacter jejuni | 40 |
| Colwellia psychrerythraea 34H | 8 |
| Desulfotalea psychrophila LSv54 | 10 |
| Methanococcus jannaschii | 85 |
| Methanopyrus kandleri | 98 |
| Pyrobaculum aerophilum | 100 |
| Pyrococcus furiosus | 100 |
| Pyrococcus horikoshii | 98 |
| Streptococcus thermophilus CNRZ1066 | 42 |
| Sulfolobus solfataricus | 80 |
| Sulfolobus acidocaldarius DSM 639 | 80 |
| Symbiobacterium thermophilum IAM14863 | 60 |
| Thermoanaerobacter tengcongensis | 75 |
| Thermobifida fusca YX | 57 |

| | |
|---|---|
| Thermococcus kodakaraensis KOD1 | 95 |
| Thermoplasma acidophilum | 59 |
| Thermoplasma volcanium | 60 |
| Thermosynechococcus elongatus | 55 |
| Thermotoga maritima | 80 |
| Escherichia coli K12 | 37 |
| Thiomicrospira crunogena XCL-2 | 25 |
| Vibrio fischeri ES114 | 28 |
| Psychrobacter arcticum 273-4 | 22 |
| Pseudomonas fluorescens Pf-5 | 32 |
| Pseudomonas putida KT2440 | 28 |
| Pseudomonas syringae phaseolicola 1448A | 26 |
| Picrophilus torridus DSM 9790 | 60 |
| Photobacterium profundum SS9 | 15 |
| Pelodictyon luteolum DSM 273 | 25 |
| Natronomonas pharaonis | 41 |
| Nanoarchaeum equitans | 82 |
| Mycobacterium avium paratuberculosis | 39 |
| Methanosarcina acetivorans | 40 |
| Methanosarcina barkeri fusaro | 35 |
| Methanosarcina mazei | 36 |
| Moorella thermoacetica ATCC 39073 | 57 |
| Methanobacterium thermoautotrophicum | 65 |
| Oceanobacillus iheyensis | 28 |
| Lactobacillus acidophilus NCFM | 41 |
| Haemophilus ducreyi 35000HP | 32 |
| Geobacillus kaustophilus HTA426 | 60 |
| Geobacter metallireducens GS-15 | 32 |
| Deinococcus geothermalis DSM 11300 | 47 |
| Chlorobium tepidum TLS | 48 |
| Carboxydothermus hydrogenoformans Z-2901 | 67 |
| Leifsonia xyli xyli CTCB0 | 29 |

| | |
|---|---|
| Clostridium acetobutylicum | 37 |
| Pyrococcus abyssi | 96 |
| Sulfolobus tokodaii | 80 |
| Streptomyces avermitilis | 27 |
| Gluconobacter oxydans 621H | 26 |
| Staphylococcus aureus aureus MRSA252 | 34 |
| Staphylococcus saprophyticus | 37 |
| Streptococcus mutans | 37 |
| Rhodopseudomonas palustris BisB18 | 30 |
| Pseudomonas aeruginosa | 40 |
| Nitrosomonas europaea | 26 |
| Pseudoalteromonas haloplanktis TAC125 | 26 |
| Shewanella denitrificans OS217 | 20 |
| Sodalis glossinidius morsitans | 28 |
| Xylella fastidiosa | 26 |
| Yersinia pseudotuberculosis IP32953 | 37 |
| Rhodospirillum rubrum ATCC 11170 | 27 |
| Magnetospirillum magneticum AMB-1 | 30 |
| Corynebacterium glutamicum ATCC 13032 Bielefeld | 33 |

---

## SUPPLEMENTARY NOTE

### Importance of selection in PAM1

In this supplementary note, we first demonstrate that the form of PAM1 matrix is predominantly determined by the genetic code, nucleotide mutation rates, and DNA composition –with little selection pressure. To do so, we plot below the MPM1 matrix computed by Nowicka et al. [3] (Fig. S1). MPM1 is computed using the empirical mutation rates for nucleotides in the *Borrelia burgdorferi* genome and a Monte Carlo algorithm that induces point mutations to achieve one-percent amino acid substitutions (same as PAM1). The two matrices are qualitatively similar, especially in the region of interest near the diagonal.

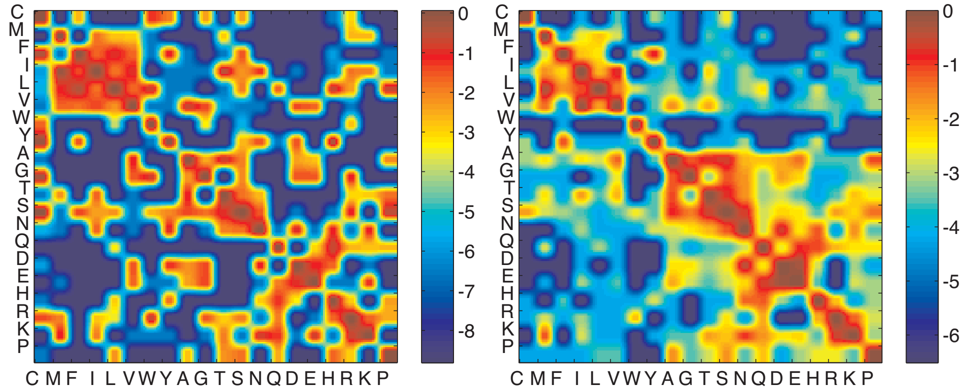

FIG. 1. Comparison of MPM1 to PAM1. (Left) MPM1 substitution matrix. Entry (i,j) is the logarithm of the probability of amino acid $i$ substituting amino acid $j$ computed using the empirical mutation rates for nucleotides in the *Borrelia burgdorferi* genome in conjunction with the genetic code [3]. (Right) PAM1 substitution matrix. Entry (i,j) is the logarithm of the probability of amino acid $i$ substituting amino acid $j$ after an evolutionary distance of one accepted point mutation for every 100 amino acids [4].

Moreover, Nowicka et al. [3] conclude that the slight differences between MPM1 and PAM1, when extended to longer evolutionary distances, indicate that amino acids with higher mutation probability are under lower selection pressure, which is consistent with our conclusion on the role of the natural composition. Computing the similarity matrix $S_{ij}$ using MPM1 instead of PAM1 for the natural and random occurrence frequencies, results in the same conclusion –that the natural frequencies enhance similarity between amino acids that are most frequency interchanged due to mutations. We present our analysis in the main text using PAM1 due to its generality, prevalent use, and intuitive association with mutation rates.

**Similarity matrix recomputed**

Herein, we establish that the improved method proposed in the first part of the paper for estimating $E_c$ from the interaction matrix and occurrence frequencies is indeed required to reach the main conclusion of the paper. To do so, we compute the similarity matrix $S_{ij}$ using the $E_c$ estimate of Eq. [2] and Eq. [3] for the natural occurrence frequencies. As demonstrated in Fig. S2, the resulting matrix no longer exhibits the intricate structures (such as a clear division by hydrophobicity and charge) seen in Fig. 5A. Furthermore, the correlations computed are mostly statistically insignificant. In fact, we needed to use 18000 subsets with highest $E_c$ (as opposed to 1000) to extract any statistically meaningful pair-wise correlations. It is also not feasible to compare random frequencies to the natural ones using this method. This confirms that the proposed scheme of diagonalization and introduction of quasi-frequencies is required for a sufficiently accurate estimate of $E_c$.

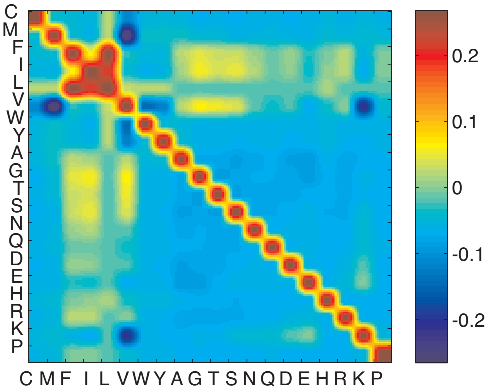

FIG. 2. Recomputed similarity matrix. Similarity matrix $S_{ij}$ computed using $E_c$ estimated from Eq. [2] and Eq. [3] of the main text. The detailed structure is no longer present and the correlations are mostly statistically insignificant.

[1] Leinonen R, Diez F-G, Binns D, Fleischmann W, Lopez R, Apweiler R (2004) UniProt archive. *Bioinformatics* 20:3236-3237.

[2] Zeldovich K-B, Berezovsky I-N, Shakhnovich E-I (2007) Protein and DNA Sequence Determinants of Thermophilic Adaptation. *PLoS Comput Biol* 3(1): e5. doi:10.1371/journal.pcbi.0030005.

[3] Nowicka A et al. (2003) Correlation between mutation pressure, selection pressure, and occurrence of amino acids. *Computational Science-ICCS-2003* 650-657.

[4] Dayhoff M-O, Schwartz R-M, Orcutt B-C (1978) A model of evolutionary change in proteins, in Atlas of Protein Sequences and Structure, ed Dayhoff M-O. (Silver Springs: Natl. Biomed. Res. Found.) 5:345-352.