Physica A 469 (2017) 676–680

Contents lists available at ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

Hidden thermodynamic information in protein amino acid mutation tables

J.C. Phillips

Department of Physics and Astronomy, Rutgers University, Piscataway, NJ, 08854, United States

HIGHLIGHTS

- Protein amino acid mutation rates are key to identifying related protein fragments.
- Most often mutated sites are either hydrophobic or hydrophilic.
- Different amino acid scales yield smoother or rougher rate-hydrophobicity curves.

ARTICLE INFO

Article history:
Received 11 July 2016
Available online 17 November 2016

Keywords:
Hydropathicity
Mutation
Scale
Blocks
Amino acid

ABSTRACT

We combine the standard 1992 20 × 20 substitution matrix based on block alignment, BLOSUM62, with the standard 1982 amino acid hydropathicity scale KD as well as the modern 2007 hydropathicity scale MZ, and compare the results. The 20-parameter KD and MZ hydropathicity scales have different thermodynamic character, corresponding to first- and second-order transitions. The KD and MZ comparisons show that the mutation rates reflect quantitative iteration of qualitative amino acid –phobic and -philic binary 2 × 10 properties that define quaternary 4 × 5 subgroups (but not quinary 5 × 4 subgroups), with the modern MZ bioinformatic scale giving much better results. The quaternary 5-mer MZ 4 × 5 subgroups are called mutons (Mu5). Among all hydropathicity scales, the MZ scale uniquely exhibits a smooth, deep mutational minimum at its center associated with alanine, glycine, the smallest amino acid, and histidine.

Protein amino acid sequences (aas) are rich in information, especially when combined with structural data. There are many Web-based tools for analyzing aas, but by far the most utilized is BLAST (**B**asic **L**ocal **A**lignment **S**earch **T**ool), which compares two given sequences, or searches for sequences similar to a given sequence. The original BLAST paper [1] was the most highly cited paper published in the 1990s. A key BLAST element is the "substitution matrix", which assigns a score for aligning any possible pair of residues, and identifies "positive" mutations between similar aas. The BLOSUM62 matrix (available online) is the default for most BLAST programs [2]. It obtains mutation rates Γ of aa pairs from protein blocks (distantly related but conserved regions), which leads to accurate homological lists of functionally similar protein blocks. The protein data base is the most extensive one in science, and extracting information from it is probably the greatest challenge in science.

Competing effects of hydrophobic and hydrophilic segments of a given protein have long been known to be the primary driving force behind the folding of protein chains into protein globules. There are secondary effects associated with longitudinal hydrogen bonding (α helices) and transverse hydrogen bonding (β strands), and even weaker charge effects, but in most proteins the dominant physico-chemical factor in a kinetic property such as aggregation [3] is hydropathic interactions.

E-mail address: jcphillips8@comcast.net.

http://dx.doi.org/10.1016/j.physa.2016.11.074
0378-4371/

-1	-1	-2	-1			3	2	0	-1
	3	-1	1				1	-1	-1
MZ		-1	1			KD		0	-1
α = CVI YM			-1			α = IVLFC			-2

Fig. 1. Two sample Mut5 α blocks, using either the MZ or KD hydropathicity scale. The integers represent logs of mutation rates, in (bit2)/2 units. Only the off-diagonal elements are shown, and the MZα rows (columns) are labeled by CVIY (VIYM). For the MZ group α, the hydropathic width is $\Psi_{MZ}(C) - \Psi_{MZ}(M)$, while for the KD group α, it is $\Psi_{KD}(I) - \Psi_{KD}(C)$.

Hydropathic interactions determine globular shapes and are manifested biochemically in many ways, which has led to many scales of aa hydropathicity Ψ. Here we will compare results obtained with the standard 1982 scale, Ψ(KD) (17k citations) [4], and the modern 2007 Ψ(MZ) scale, based on fractals and self-organized criticality [5]. The KD scale is related to first-order effects (unfolding of globular proteins from water to air), while the MZ scale describes second-order conformational changes in globular surface differential geometry [3]. Our analysis of the BLOSUM62 matrix will enable us to decide whether block homologies are primarily first- or second-order thermodynamically. We also are looking for non-equilibrium effects associated with mutations: how are these represented by thermodynamic scales?

There is already a large literature on general aspects of biological evolution and statistical physics [6], which aim to go beyond phylogenetic trees based on point mutations (much less effective than block similarities [2]). Explicit applications help to bring these general considerations into sharper perspective. The biomedically important area of rapidly mutating viruses and their vaccines is best quantified using epitopes [7,8], which are similar to but still different from blocks, which are best suited to describing self-sustaining proteins [9].

Mutational rates can be used to compare two aspects of different hydropathicity scales Ψ(aa), their discrete hierarchical ordering of 20 amino acids, with 20! possible orderings, or the continuum spacings between ordered amino acids. The MZ and KD orderings [5] can be divided into groups (hydropathic scale blocks), with the two obvious choices for replacing binary $2 \times 10 = 2 \times (2 \times 5)$ by $4 \times 5 = (2 \times 2) \times 5$, or quaternary 4×5aa blocks (denoted by Mut5), or quinary 5×4aa blocks (Mut4). The most hydrophobic Mut5α blocks, extracted from the BLOSUM62 matrix, are shown for the MZ and KD scales in Fig. 1. These groups could display the tendency of amino acids to mutate into other amino acids within their subgroup with similar hydropathicity.

One often sees qualitative comparisons of mutation rates of hydro (phobic, philic) aa, but with Mut groups one can make quantitative comparisons. One averages the mutational off-diagonal group matrix elements, and compares those averages with the hydropathic width of each group (defined as Ψ(first aa) $-$ Ψ(last aa)). The wider the subgroup, the more Ψ phase space is available for internal mutations. This idea can be tested at the simple 2×10 hydrophobic/hydrophilic level, for $\Omega = (\text{Mut4}\alpha + \text{Mut4}\beta) - (\text{Mut4}\gamma + \text{Mut4}\delta)$. With the MZ scale $\Omega = -1.2$ (as one might have expected, exposed hydrophilic aa mutate much more often than buried hydrophobic aa), but with the KD scale $\Omega = 0.2$ (unsatisfactory).

This binary phobic/philic test is coarse: what happens when we calculate the Mut5 correlations of average mutation rates with average hydropathic widths? The results are $R = 0.93$ (MZ) and 0.86 (KD), both very successful, but MZ is even more successful. When we repeat these steps with the Mut4 groups, we obtain weak and inconsistent results: $R = 0.3$ (MZ) and -0.3 (KD). The 4×5 Mut5 partition is an iteration of the 2×10 phobic/philic partition, which explains its block success, as well as the failure of the non-iterated Mut4 blocks.

The symmetry of BLOSUM mutation rates suggests that we examine successive waves (generations) of mutation rates, corresponding to diagonal strings parallel to the principal (unmutated) matrix diagonal. Again we reduce noise by looking at average rates $\Phi(N)$ of groups, but now the groupings $\Lambda(N, N+\kappa)$ are averaged over waves based on the hydropathically ordered aa sequences (in matrix terms, Γ elements $(N, N+\kappa)$ are averaged over κ, from $\kappa = 1$ to κ_{max}). The results using the MD ordering are shown in Fig. 2, and the KD ordering in Fig. 3, and discussed in those captions.

A striking feature of the mutational waves is the smoothness of the MZ $\Phi(N)$ groups (Fig. 2) compared to the KD $\Phi(N)$ groups (Fig. 3). Here we define roughness as the 20-aa average over N of $(\Phi(N) - \Phi(N+1))^2$. This roughness is presumably a measure of the thermodynamic noise of block mutation rates, averaged over thousands of proteins. As shown in Fig. 4, this noise is almost the same for the MZ and KD scales for $\kappa_{max} \geq 10$, but for $\kappa_{max} = 5$, MZ is 35% smoother, presumably reflecting a greater information content using the thermodynamic fractal MZ conformational scale, compared to the unfolding KD scale. Note that it has been found that ~4 mutations within epitope A or B, between the old vaccine target strain and the currently dominating circulating strain, are enough to render the H3N2 vaccine ineffective [7–9]. This is a high level of internal consistency between two widely separated methods.

The difference between Figs. 2 and 3 suggests a simple criterion for optimal hierarchical ordering of hydrophobicities in the many (~126) scales which were suggested in the classic period [10]. According to Fig. 2, the deep hydroneutral minimum in mutation rates with the MZ scale is associated with alanine (A), glycine (G), the smallest amino acid, and histidine (H). Table I of Ref. [5] shows that none of the older scales places all three of these amino acids at its center. In terms of RMS deviations from the average value of each scale, the off-center differences are 7 times or more larger for the other scales than for the MZ scale. Note that these differences are not much larger than the quoted error bars in the original work; the fractal MZ scale, based on solvent accessible areas, is more accurate in principle, but it also benefits from its post-2000 bioinformatic survey of more than 5000 segmental structures.

The stability of histidine makes it the central and most conserved element of many catalytic triads [11], the most studied examples being Serine–Histidine–Aspartate (chymotrypsin) and Cysteine–Histidine–Aspartate. Catalytic triads form