Decoding the Design Principles and the

Chemical Logic of Amino Acids

Supplementary Information

1. **Supplementary notes on the assignment of the four unique chemical properties to amino acids** (subscripts in the last column of Table 1 in the main text) .

Exactly 10 amino acids {E, I, K, L, M, P, Q, R, T, V} possess sp^3 hybridized γ carbon atom (property I denoted by symbol **g**). The number of unique sp^3 hybridized carbon atoms to which the sp^3 γ carbon is attached is counted as the number of times **g** occurs. This is shown as subscript to **g** in the last column of Table 1 in the main text. Thus the following assignments can be readily made: $\mathbf{g_1}$ for E, $\mathbf{g_2}$ for I, $\mathbf{g_3}$ for L, $\mathbf{g_1}$ for M, $\mathbf{g_2}$ for P, $\mathbf{g_1}$ for Q, $\mathbf{g_2}$ for R, $\mathbf{g_1}$ for T and $\mathbf{g_1}$ for V. $\mathbf{g_1}$ for K follows from rule 2(a). Other amino acids are assigned $\mathbf{g_0}$.

Again, exactly 10 amino acids {A, C, D, G, I, M, N, S, T, V} lack a δ carbon (property III denoted by symbol **s**). The number of carbon atoms short of the δ carbon in the side chain is counted as the number of times **s** occurs. This is shown as subscript to **s** in the last column of Table I. Glycine does not have a c_β, c_γ and c_δ. Thus it is assigned $\mathbf{s_3}$. Alanine does not have c_γ and c_δ . Thus it is assigned $\mathbf{s_2}$. Thus the assignments are: $\mathbf{s_2}$ for A, $\mathbf{s_2}$ for C, $\mathbf{s_1}$ for D, $\mathbf{s_3}$ for G, $\mathbf{s_1}$ for I, $\mathbf{s_1}$ for M, $\mathbf{s_1}$ for N, $\mathbf{s_1}$ for T, $\mathbf{s_2}$ for V. $\mathbf{s_1}$ for S follows from rule 2(a). Other amino acids are assigned $\mathbf{s_0}$.

It may be noted that by virtue of the above assignments and properties of side chains indicated as satisfied in Table 1 and given that the properties present in any amino acid i.e. the subscripts in the last column of Table 1 must add up to 3, the

assignments are nearly complete except for H and Y. Thus the **d** and **l** assignments for amino acids are as follows: $\mathbf{d_0\ l_1}$ for A; $\mathbf{d_1\ l_0}$ for C; $\mathbf{d_0\ l_2}$ for D, $\mathbf{d_0\ l_2}$ for E, $\mathbf{d_0\ l_3}$ for F, $\mathbf{d_0\ l_0}$ for G, $\mathbf{d_2\ l_1}$ for H, $\mathbf{d_0\ l_0}$ for I, $\mathbf{d_1\ l_1}$ for K, $\mathbf{d_0\ l_0}$ for L, $\mathbf{d_0\ l_1}$ for M, $\mathbf{d_2\ l_0}$ for N, $\mathbf{d_0\ l_1}$ for P, $\mathbf{d_2\ l_0}$ for Q, $\mathbf{d_1\ l_0}$ for R, $\mathbf{d_1\ l_1}$ for S, $\mathbf{d_1\ l_0}$ for T, $\mathbf{d_0\ l_0}$ for V, $\mathbf{d_3\ l_0}$ for W, $\mathbf{d_1\ l_2}$ for Y. An interpretation of hydrogen bond donor (**d**) assignments quite interestingly suggests that it is not only the presence of a hydrgen bond that is important but also the geometrical arrangement in which a hydrogen bond donor atom is presented.

2. **Supplementary notes on the protein and genome sequence analyses**

All the available Swissprot sequences are considered for the analysis here except those which are annotated as hypothetical or putative or contain less than 25 amino acids. Excess values reported in Table 2 of main text are relative to random sequences. The word length is taken as 11 amino acids. Excess $g = (g - g_{random})$; $gl=(g^2+l^2)^{1/2}$ etc. are computed for each word using the assignments in Table 1 after converting them into unit vectors and summing the value of each property (g,d,s,l) over all words for each sequence and the property reported for a sequence of length 100 amino acids. Value of g, d, s, l for a random sequence containing 100 amino acids is 33.8 when the g,d,s,l values (last column, Table 1) for each amino acid are converted into unit vectors. By random sequences is meant, all amino acids occur with equal probability at every location along the sequence.

To identify whether a query nucletotide sequence is likely to code for a protein (results in Table 3), the sequence is converted into amino acid space using the genetic code. The polypeptide sequence thus generated is treated as a collection of

words. The chemical properties of amino acids in the query sequence are summed for each word using the values provided in Table 1 after converting the values for each amino acid into a unit vector and traversing along the whole sequence word by word. The sequence is examined for its compliance with the observation of excess of 'g', 's' and low 'd', 'l' together with the standard deviations and the composite properties, 'gs-dl >0' etc. (Table 2). In addition, for results in column 4 of table 3, if a stop codon occurs within an interval of 25 codons from the starting methionine (ATG), the sequence is considered as a nongene.

3. **Supplementary notes on the statistical indices evaluated for data presented in Table 3 of the main text.**

True Positives (TP)= genes identified as genes;

True Negatives= non-genes identified as non-genes;

False Positives= non-genes identified as genes;

False Negatives= genes identified as non-genes;

Number of Actual Positives (AP) = TP+FN

Number of Actual Negatives (AN) = FP+TN

Predicted Number of Positives (PP) = TP+FP

Predicted number of Negatives (PN) = TN+FN

Sensitivity (SS) = TP / AP

Specificity (SP) = TP / PP

Correlation Coefficient (CC) = $(TP.TN - FP.FN) / (AN.PP.AP.PN)^{1/2}$