A hydrophobic proclivity index for protein alignments
[version 2; peer review: 1 approved with reservations, 1 not approved]

David Cavanaugh¹, Krishnan Chittur²

¹Benchmark Electronics, Huntsville, AL, USA
²Chemical Engineering Department, University of Alabama Huntsville, Huntsville, AL, USA

Abstract
Sequence alignment algorithms are fundamental to modern bioinformatics. Sequence alignments are widely used in diverse applications such as phylogenetic analysis, database searches for related sequences to aid identification of unknown protein domain structures and classification of proteins and protein domains. Additionally, alignment algorithms are integral to the location of related proteins to secure understanding of unknown protein functions, to suggest the folded structure of proteins of unknown structure from location of homologous proteins and/or by locating homologous domains of known 3D structure. For proteins, alignment algorithms depend on information about amino acid substitutions that allows for matching sequences that are similar, but not exact. When primary sequence percent identity falls below about 25%, algorithms often fail to identify proteins that may have similar 3D structure. We have created a hydrophobicity scale and a matching dynamic programming algorithm called TMATCH (preprint report) that is able to match proteins with remote homologs with similar secondary/tertiary structure, even with very low primary sequence matches. In this paper, we describe how we arrived at the hydrophobic scale, how it provides much more information than percent identity matches and some of the implications for better alignments and understanding protein structure.

Keywords
Sequence alignment algorithms, hydrophobicity scale, protein homologs, TMATCH

Any reports and responses or comments on the article can be found at the end of the article.
Proteins with low sequence identities, possessing statistically insignificant similarities by conventional measures, but having similar secondary/tertiary structures, which would not be identified as statistically significant by other methods such as FASTA and Smith-Waterman can be identified as homologous using a new alignment algorithm (manuscript in preparation) through the enhanced information content of our hydrophobicity proclivity scale.

Hydrophobicity scales (also often called metrics) as understood in the literature are generally divided into four categories, derived from

- Experimental physio-chemical data
- Log of a partition coefficient derived from protein structure (e.g. Fraction amino-acids inside vs. outside, fraction amino-acids in contact with water vs. completely buried, etc.)
- Amino-acid mutation/substitution rates and
- Participation rates/probabilities of occurrence in folded protein secondary structure

There are a large number and myriad types of scales that appear in the literature starting from the 1960’s through to the present with a fair amount of variation amongst these scales. The correlation between some of the hydrophobicity scales can be best understood as that derived from the energy of interaction between amino-acids and water or the energetics of partition of amino-acids from water as the reference state and some other environment such as a non-polar solvent or the interior of a folded protein. Hydrophobicity can thus be joined within a single, unified, conceptual framework.

Our hydrophobic index is the result of an extensive mining of the literature about proteins and amino acid scales/metrics in different environments. Almost all hydrophobicity scales reflect in some way a measure of the energetics of transfer of an amino-acid (or proteins) from one solvent environment (water) to another (folded protein or multiple protein assembly). During our data mining and analysis, three hydrophobicity metrics emerged as the most appropriate since we could relate those scales to multiple fundamental properties of the 20 natural amino acids using multi-variate statistical procedures, thermodynamics and biophysical chemistry considerations.

Hydrophobicity scales reflect different physical properties of amino-acids, such as metrics derived from aminocid partitioning patterns (e.g. from the hydrophobic core to the exterior of proteins) or log of partition ratios between water and organic solvents. We found, as widely suggested in the literature, that the free energy of transfer from water to octanol turns out to be a good proxy for the hydrophobic core environment of folded proteins.

Our hydrophobicity proclivity scale was deployed as part of a new protein alignment algorithm TMATCH, which is our adaptation of the Needleman-Wunsch dynamic programming alignment algorithm that uses TMATCH has adapted the
Needleman-Wunsch fundamental dynamic programming alignment algorithm. In TMATCH, the local alignment score reinforces favorable diagonal sequences that are paired with fixed gap opening penalties. The TMATCH algorithm is especially designed to take advantage of the extra information available within our hydrophobicity scale to detect homologies, as opposed to the probabilities derived from raw percent identities.

Low sequence identities, possessing statistically insignificant local alignment similarities using methods such as FASTA and BLAST and using conventional percent identity measures, but having similar secondary/tertiary structures, can be identified using our TMATCH algorithm and our amino-acid hydrophobicity proclivity scale.

The TMATCH\(^{5,6}\) algorithm uses a fixed gap penalty and therefore abandons the notion of an affine gap penalty (based upon a linear weighting function), which is problematical as there is no deep, underlying theoretical construct for choosing a specific affine gap penalty function, which derives from statistical theory and/or from protein function/structure. The TMATCH algorithm uses the fact that local pair-wise sequences of high homology result in diagonal (upper left to lower right) traces in dotmatrix/dot-plot algorithm. When these local pairwise diagonal traces exist in the optimal/near-optimal alignment catchment basin, which is defined as an area about the dot-matrix major diagonal, they will contribute to and be included within the global/near global alignment traces within the dot-plot/dot-matrix.

TMATCH\(^{5,6}\) captures these dot-plot/dot-matrix algorithmic properties by introducing the notion of score “rewards” for favorable (e.g. local alignment score optimization) cell-cell diagonal transitions and score “punishments” for unfavorable cell-cell diagonal transitions. Fixed gap penalties are assessed for horizontal or vertical cell-cell transitions. Pairwise comparisons of amino-acids in the alignment are done with hydrophobicity proclivity scale. In addition to the diagonal score for a given table cell, the traditional dynamic programing score algorithm is used for calculating entry from the left or above subject to the fixed gap opening penalty.

The similarity calculated for two aligned proteins are based upon the percent hydrophobic fuzzy match similarity of two protein sequences being aligned. The average number of pairwise fuzzy matches in the alignment allow for the computation of a statistical relationship test without having to actually extract the alignments, thereby saving the associated computational overhead.

As an alignment search algorithm, we have carefully benchmarked the TMATCH\(^{5,6}\) algorithm performance against several protein families and achieved very good results as compared to FASTA, BLAST and PSI BLAST. Specifically, we have worked with Glutathione proteins, G proteins, Rhodopsin proteins, Tryptophan like Serine proteases and DNA Polymerase B enzymes. The latter two protein families are especially challenging due to having low percent sequence identities, yet highly conserved tertiary structures. For the DNA Polymerase B and Tryptophan like Serine Protease families, we have seeded the protein sequence cohorts with proteins that are not within the families, but will come somewhat close to these two families with normal alignment algorithms and/or cause alignment algorithms to have pathological behaviors. The TMATCH search alignment algorithm did very well in these trials.

Methods

Approach

Our primary and starting premise was to look for clustering, single/multiple linear patterns or non-linear patterns between all amino-acid scales within our database as seen in the scatter plot of any two-given amino-acid scales. Extensive cross correlation with the primary and derived (from primary properties or experimental scales) metrics was done using regression modeling in order to recover the best and most meaningful hydrophobicity metrics. Through this extensive analysis, we were able to identify patterns and arrive at metrics describing amino acid properties. We derived a number of additional metrics by differentiating metrics that were intrinsic as opposed to extrinsic, as understood in thermodynamics.

Along the way, we found that multiple linear series occurring in scatter plots between many pairs of amino-acid properties, which was a striking pattern. Many of these multi-family linear series ranged between clearly discernible to very high quality; the latter end of the range being what we concentrated on. We also cross checked to see if sister AA property scales in the same class such as hydrophobicity resulted in similar patterns, such as which we see in Figure 1 and Figure 2 and in Table 2. Once the linear/non-linear (single and multi-family) patterns were found a detailed review of these patterns were made to find underlying physico-chemical and biological reasons for these patterns as well as statistical generalizations. The most relevant scatter plots were selected based upon their quality (visual and linear/non-linear regression) and explanatory power for protein structure and function, and reliability/specificity for protein alignments.

We relied on several different sources for our analysis. For data on amino-acid surface areas, we used Rose et al.\(^7\). Amino acid mass information was obtained using the AAIINDEX accession number #FASG760101\(^8\). Amino acid volume data was obtained from Creighton\(^9\). Amino acid absolute entropy of formation was from the AAIINDEX database using accession number #HUTJ700102\(^9\). We arrived at our hydrophobicity scale after exhaustive analysis which included numerous scatter plots and the running of a number of multiple regressions. The question we were trying to answer was - What was the best hydrophobicity scale, or combination of scales, that best represented the role of the different amino acids in proteins?

Hydrophobicity

We started by first collecting many hydrophobicity indices and physico-chemical indices from the literature and scatter plotted/regressed the hydrophobicity indices against each other, and the harvested physico-chemical properties and their
Figure 1. Normalized average of several hydrophobic scales with Solvent Exposed area. Note that the Y scatter tends to go up with polar and ionic amino acids.

Figure 2. Hydrophobic Proclivities versus Structure F & P Gtow. Tryptophan and Proline diverge from the regression line and this can be explained by the ring structure of the secondary group being relatively wide, which impacts the surface tension of the corresponding aqueous clathrate and the stearic packing effects of these two residues in a folded protein.

derived intrinsic properties of amino acids. For example when a hydrophobic scale is plotted against the ratio of the surface area per specific volume (volume/molecular weight) for each amino acid we get a scatter plot with a distinct pattern. In such a scatter
plot, we can identify one or more sets of clusters/linear clusters of amino acids, each set of which is considered to be a "property class".

We were able to find three hydrophobicity scales that were the most robust from the regression cross correlation study. The hydrophobicity proclivity scale that we report in the present paper are the normalized average of three normalized scales\(^2\)–\(^4\)

We created a normalized average of the three key hydrophobicity scales (The index \(i=1\) is from Tang\(^2\), index \(i=2\) is from Neumaier\(^1\) and the index \(i=3\) is from the average of the collected scales in Juretic\(^3\)). This normalized average of three scales provides a reasonably unbiased estimate of the "true" average hydrophobicity relationship amongst the 20 amino-acids (index \(j\), from 1 to 20)

\[
H_i(\bar{b}) = \frac{H_i - \min(H_i)}{\max(H_i) - \min(H_i)}
\]

\[
H_i(\bar{b}) = \frac{H_{i1} + H_{i2} + H_{i3}}{3.0}
\]

The hydrophobicity scale as calculated using Equation 2 using the scales published in 2–4 has a number of interesting relationships with key physico-chemical properties of the amino-acids in proteins. For example, this normalized average of these three best hydrophobicity metrics possesses statistically significant linear correlation with many other reliable hydrophobicity metrics derived from multiple literature hydrophobicity scales.

Additional amino-acid property indices
We selected 49 different amino acid properties from a large data set of more than several hundred properties on the Japanese genome net AAINDEX database and from other literature sources. Our objective was to span a wide range of properties that could be used to model the most important underlying central tendencies of the amino acid properties when they are in different proteins and in different contexts. We were interested in determining how we could reduce the rather large number of studies of amino acids in different environments into a core set of properties/values that could best represent all 20 natural amino acids. Our goals for selecting amino-acid property scales from the AAINDEX database was to obtain a balance of molecular properties, statistical properties describing average secondary structure proclivities and average amino-acid fraction compositions, experimentally measured HPLC retention times, and experimentally measured amino-acid bulk properties. Twelve of the amino acid property scales were selected to represent measures of average amino-acid fractional burial in folded proteins and/or most popular/widely used literature hydrophobicity scales. The 49 AA properties are represented by:

- 2 sequence frequency scales
- 6 secondary structure propensity scales
- 8 hydrophobicity scales
- 4 free energy scales (in water, protein folding/unfolding) n-Octanol to water (Gtow) scale
- 8 HPLC retention time scales
- 4 probabilities of an AA inside a folded protein core or on the outside
- 7 molecular property scales
- 10 physical (experimentally measured) property scales

There were approximately 175 primary amino-acid property scales of particular interest harvested from the literature to represent the broad range of physico-chemical properties to be found in the literature. We down selected for 49 amino-acid scales used for the multi-variate analysis (PCA, nmMDS, ANOPA) portion of our study, of which around 25 were derivative scales prepared from ratios of fundamental molecular properties such as volume and mass, which would represent intrinsic versus extrinsic scales in the thermodynamical sense. About another 10–15 scales represent averages of literature reported amino-acid property scales. Note that the scales selected represent a very wide range of amino-acid properties including secondary structure statistical propensity scales, fundamental molecular properties like surface area/mass/volume, bulk properties such as index of refraction/melting point/pK-C, HPLC retention times, free energy in water, solvent/water partitioning, average fraction occurring in proteins, average amino-acid burial in proteins/amino-acid exposure with water in folded proteins, hydrophobicity scales, NMR parameters, RF parameters, several literature Principal Component Analysis/Factor Analysis parametric scales and other miscellaneous property scales.

Multi-variate analysis
The analytical method we used for this data is called the analysis of patterns (ANOPA). In some ways, the ANOPA analysis process resembles "factor analysis."

ANOPA is a pattern projection method. The ANOPA procedure projects n-space pattern point/vectors into a 3-dimensional object which is a cylinder. The axis of the cylinder, called a relation vector, is formed from the pattern point centroid (averages of each AA property for all points) to an out-group average point (averages of each AA property for a pair of selected points). The out-group pair are selected on the basis of a histogram of the pattern point Euclidean distances from the pattern point centroid. Two pattern point distances are calculated with respect to the relation vector from a projection of each point onto the relation vector yielding the distance along the relation vector and the distance from the relation vector. The angle of rotation of each pattern point projection about the relation vector is calculated. Thus, a cylindrical coordinate system is formed which is then converted into rectilinear coordinates. The X prime and Y prime rectilinear coordinates are formed from each pattern point’s pattern projection distance times the cosine and sine respectively of the pattern projection angle of rotation. The Z prime coordinate is simply the pattern projection intersection distance along the relation vector.

We confirmed the results of the ANOPA analysis through a Principal Component Analysis (PCA) analysis of the same 49 amino-acid property scales and calculation of correlation
correlation coefficients between the 3 axes of each of the PCA analysis and ANOPA analysis. To even add more confidence to confirm the ANOPA findings we also run a non-metric Multiple Dimensional Scaling (nmMDS) analysis on the same 49 amino-acid property scales. These latter two analysis were conducted with the software package PAleontology STatistics - PAST\(^{11,12}\). The ANOPA calculations were performed in the Microsoft Excel 2010 package with the established ANOPA equations\(^{13}\) and cross checked with a de novo Python software implementation.

**Correlation coefficients and statistical significance testing**

The correlation coefficients, simple linear regressions and the Multi-Linear Regression (MLR) analysis were computed using the statistics functions and statistics data package of the Microsoft Excel 2010 package. There are two statistical tests run on the MLR regressions involving the Property Class (PC) scales (for example Table 3). The first is the F test that comes with the stock standard Microsoft Excel 2010 statistics package MLR add in software. The second statistical test used is a Students T test, where the T statistic is a standard transform of the correlation coefficient that possesses a Students T distribution with the null hypothesis that \(R = 0\) and there is no statistically significant relationship. The correlation coefficient T test number of degrees of freedom (20) is reduced by 2 for each X variable coefficient in the regression and by 1 for the Y axis intercept.

\[
T(R) = \frac{N-2}{1-R^2} \tag{3}
\]

Where \(N=20\) for the number of amino-acids. The conservative degrees of freedom for the 8 Property Class scales (described in the results section and found in Table 3) are calculated as 20 less the loss of degrees of freedoms for the MLR regression, where the degrees of freedom are 20 less two times eight (average and standard deviation for each PC) and less 1 for the regression intercept, or \(d f = 20 (2^8 + 1)\). The critical value of the T (df =3) statistic is 2.3604, yielding a P value of 4.968\%, which we round to the customary statistically significant Alpha tail (df =3) statistic is 2.3604, yielding a P value of 4.968\%, which we round to the customary statistically significant Alpha tail area P =5\%. Correspondingly, the minimum \(R^2\) to have statistical significance in the present study is 0.65, which is a correlation coefficient \(R = 0.8062\).

**Notes on Multiple Linear Regression (MLR)**

The argument we present herein is not that the MLR regression is superior to that of the binary regression, because of higher correlation coefficients, although possibly that may be true, but rather that the amino-acid property class MLR represents more information because more properties represent more of the behaviors of amino-acids in a larger series of contexts and behaviors in real proteins. Our argument also contends that by having a MLR with the 8 property class scales (Table 3), for example, reflects more contexts because the amino-acids partition into sub-sets and that these different context sub-sets join to determine amino-acid behavior in protein folding, interactions with water, interactions with membranes, secondary structure and electronic behaviors associated with AA-AA interactions. Moreover, we argue that within any given regression that the higher the correlation coefficient, the stronger the evidence for the superiority of a given hydrophobicity scale within that regression criterion. Top performance of a hydrophobicity scale within several regression relationship criterion’s based upon different property information leads to a performance/robustness conclusion based upon the consilience of the evidence. We also note that the practice of encoding of attribute, qualitative or state data as integers for the purposes of statistical analysis is widespread and fruitfully used in engineering and science, especially within the enterprise of taxonomy in Biology.

We use the coefficient of determination \(R^2\) rather than the correlation coefficient because it is a much more conservative statistic than the correlation coefficient \(R\) and can meaningfully be interpreted as the percent linearity between the dependent variable and the X independent variable(s) used in the regression. Concerns might be raised concerning two points regarding the property class MLR, which are the possibility of inter-correlation between X variables and the loss of degrees of freedom in the regression statistics (over-fitting). We deal with these concerns in several fashions. First off, we calculate a T test significance only for the regression itself and not for any individual X variable regression coefficient since the Y variable regression performance itself is what is being measured (see for example Table 11). Secondly, we introduce a correlation coefficient T test as described above. Thirdly, we establish general and specific regression observations as verified over many regressions.

**Statistical significance of the regression vs. statistical significance of the regression variables/factors**

We allow for some dimensionality reduction due to inter-X pair correlation as long as the dimensionality reduction (i.e. percent reduction in the number of states with discrete variables) is not large and new information is being brought forth between each pair of X variables. Property class two partitions into two linear subgroups on the basis of polar and non-polar amino-acids. Property class four partitions into 4 linear subgroups on the basis of size, relative amounts of non-polar area and the presence/absence of a relatively strong or weak polar group. While property class 4 does reflect the polar/non-polar distinction to some extent, it does not reflect a strong and clear partition of AA by polarity or non-polarity, therefore new information is introduced by the addition of property class 4. Property classes 2 and 4 are highly correlated by classical measures and have a Kendall tau coefficient of 0.82 and a Pearson correlation coefficient of 0.86. We note that there are 2^4 =16 possible combinations of states between property classes 2 and 4, but in fact there are actually 5 states formed thereby giving a dimensionality reduction of 1.55/8 =37.5\%, which we consider to be an acceptable price for the new information. Where there is no new information added by inclusion of a X variable we find that the magnitude of the correlation coefficients are reduced and even more significantly reduced by the squaring process to get the coefficient of determination. Furthermore, we demand that any set of property class vectors produce large significant coefficient of determinations with many disparate AA property scales in our database. The latter criterion produces a very, very high degree of statistical confidence that the collection of property class scales can serve as a highly robust basis set for MLR vector regressions that
can evaluate the robustness/significance of many individual AA property scales. There are no other statistically significant inter-property class correlations (see Table 9).

Another key point is that the MLR X’s are not single variables, but rather are column vectors with each column vector having values for each amino-acid. In this context, equate the concept of a scale with the concept of a column vector. Having made this distinction, the concept of a Multi-Linear Regression (MLR) is still valid and simply represents a broader mathematical context. Our use of the MLR methodology with the 8 property class scales (Table 3) we develop is to evaluate a correlation relationship as whole entity, so the entire MLR correlation is what is used/important, and the statistical significance of individual variable regression coefficients is not meaningful in our context as we do not seek to assess the statistical significance of individual vectors. If the MLR method was being used to fit a variable for the purpose of extrapolating new Y values with scale values outside of the basis set, such as with new amino-acids, then the statistical significance of individual variable coefficients would be germane.

Data point clustering as a way of defining “fuzzy” sub-sets assigned ordinal numbers as regression variables

When doing AA property scale pair scatter plots between each property scale selected for our database we often find clustering behavior which indicates that the amino-acids are partitioning into distinct sub-sets and each amino-acid can be assigned a numerical value for the ordinal number of the sub-set into which it partitions. The clustering relationships uncovered for any particular pair of two AA property scales is generally driven by molecular geometry, secondary group geometry, numbers and types of atoms/inter-atomic bonding in the secondary group, molecular size, molecular mass, molecular volume, molecular surface area and entropy distribution about the molecule. No given pair of AA property scale relationships represents the full range of these molecular properties and the nature of the interaction with water and cellular membranes.

The key point though is that there are distinct sets and sub-sets, which when numerically encoded can jointly describe each amino-acid as a row vector of some dimensionality M. We have found through extensive analysis that M=8 is a very good size. Within each property pair scatter-plot where clustering occurs, we can have different patterns such as unstructured clusters, multiple quasi-parallel linear clusters or multiple linear clusters that intersect at some point (often at Glycine or Alanine). Karplus\(^5\) noted this pattern, although not as extensively as what we report. Generally, there is a geometric ordering that allows the assignment of a meaningful ordinal number. To reiterate, a property class scale (column vector) represents a relationship between property scales and the values assigned within the scale to each amino acid are their respective sub-set/cluster ordinal number. Property class scales assembled in this way can be used as one of the basis set vectors to be used to evaluate the performance and reliability of any individual amino-acid property scale.

**ANalysis Of Patterns - ANOPA equations**

We report herein a recast of the ANOPA linear algebra vector equations\(^5\) in the as implemented form. The ANOPA equations are cast in a quasi-software convention. We need to note that the three property class vectors defined using the 3D ANOPA coordinate system planes represents the correlation based dimensionality reduction from a pattern space of dimension 49 to a pattern space of dimension 3.

- Pattern vectors (points), N objects (row vectors, amino-acids), M properties (dimensions)
- Resulting object property array/table O(M,N)
- Table Row index \(i\) =1-N (can imply column vector of results)
- Table Column index \(j\) =1-M (can imply row vector of results)
- Vector without an index implies the vector as a whole entity
- Vector norm of any vector G is symbolized by |G|
- Normalized object table \(O_{i,j} = (O(i,j)-\text{MIN}(O(i,j)))/(\text{MAX}(O(i,j)) - \text{MIN}(O(i,j)))\)
- Group Centroid/Average Ca(j) =((Sum \[i =1-N\] - >No(i,j))/N
- Dimension Delta\(^2\) \(Dd(i,j) = (\text{Range}[i =1-N] - >(\text{Range}[j=1-M]) > (\text{No}(i,j) - \text{Ca}(j))^2\)
- Distance of object pattern points to group centroid \(A0(i) =((\text{Sum}[j =1-M] - >d(i,j))^2)^{0.5}\)
- Select two objects (amino-acids a + b) \(\rightarrow \text{No}(a,j)\) and \(\text{No}(b,j)\) to calculate an outgroup average per the selection rules from the histogram of pattern point \(A0(i)\) distances to the pattern set centroid
- \(A0(i)\) distances to centroid \(Oa(j) = (\text{No}(a,j) + \text{No}(b,j))/2\)
- Out group to centroid delta \(U0(j) = (\text{Oa}(j) - \text{Ca}(j))\)
- Normalized dist along relation vector \(T0(i) = ((\text{Sum}[j =1-M] - >(\text{No}(i,j) - \text{Ca}(j))^2)*U0(j)/|U0(j)|)^{0.5}\)
- Origin projection \(T0\) type distance \(Ot = ((\text{Sum}[j =1-M] - > (\text{No}(i,j) - \text{Ca}(j)))*U0(j)/|U0(j)|)^{0.5}\)
- Origin projection reference vector \(Ur(j) = (\text{Ot}^*U0(j)/\text{Ca}(J))\)
- Pattern point dist from relation vector \(D2(i) = ((\text{Sum}[j =1-M] - > (\text{Ca}(j) - \text{No}(i,j) + T0(i))^2)/|U0(j)|)^{0.5}\)
- Cosine of pattern projection vector to relation vector with origin ref vector \(Ct(i) = ((\text{Sum}[j =1-M] - > (\text{Ca}(i,j) - \text{No}(i,j) + T0(i)*U0(i)))\)*Ur(j)/(|Ur|*D2(D))\)
- Angle of rotation about relation vector \(Ar(i) = \text{ARC-COS}(Ct(i))\)
• Distance along relation vector \( Z'(i) = T(i)^*|U(0)| \)
• Rectilinear X from D2 and Ar: \( X'(i) = D(2)i^*COS(Ar(i)) \)
• Rectilinear Y from D2 and Ar: \( Y'(i) = D(2)i^*SIN(Ar(i)) \)

**Results**

**Scatter-plots of literature hydrophobicity scales**

Hydrophobicity scales are typically derived from a measure of the probability that a particular residue will be buried in the core of the protein, away from water. What confounds these calculations is the fact that in most proteins, many of the hydrophobic residues are still exposed to the water (solvent). It is often not clear on how to treat residues that have properties intermediate between hard core hydrophobic and polar residues. The size of the residues and difference between aliphatic and aromatic residues also pose some difficulty in the calculation of a hydrophobicity scale. Calculations involving cysteine residues add additional complexity in that some of those residues may be involved in providing proteins structural stability through formation of disulfide bonds. Thus, calculation of contributions to any hydrophobicity index through analysis of where specific residues are in a given protein has been complicated and contributed to the scatter we see in the data.

We demonstrate this scatter (in this case level sensitivity) by examining the normalized average of several popular hydrophobicity scales (26-29) versus the probability of an amino-acid solvent-exposed area (SEA) (22,23) greater than 30 square Angstroms (shown in Figure 1). A good contrasting example scale has been published in 1 and is derived from an analysis of 28 literature hydrophobicity metrics, which when compared with our normalized average of three hydrophobicity scales that forms our hydrophobicity proclivity scale, possesses a strongly linear relationship \((R^2 = 0.959)\) illustrating a relationship with relatively little data scatter.

Figure 1 shows that there is indeed a relationship between the average of five hydrophobicity scales described above and whether or not a particular amino acid is within a protein core or exposed on the surface. We see one tight grouping of amino acids in the figure (I, F, V, L, M, W, A and G) and two loose groupings that include P, T, S, Y, H and N, Q, E, D, K and R. The group at the top right \((N, Q, E, D, K \text{ and } R)\) include amino acids that are ionic/strongly polar and the central group of amino acids are of intermediate polarity. The tight group of amino acids are primarily amino acids with hydrophobic residues. As we go from the very hydrophobic group to the less hydrophobic group (from the lower left to the top right) the scatter goes up. This scatter is indicative of the increase in water amino acid interaction and of the difficulty of accurately calculating the contribution of any particular residue.

**Delta Gibb's free energy of transfer between water and N-Octanol (\( \text{dGtow} \))**

In Figure 2 we show a scatter plot of our amino-acid hydrophobicity proclivities against the popular Fauchere & Pliska free energy of amino-acid transfer from n-Octanol to water (Gtow) scale (25,26). It is common in the literature to see n-Octanol used as a proxy for the typical hydrophobic core of folded globular proteins, consequently the Gtow scale has been widely used as a measure of hydrophobicity. As can be seen in Figure 2 the correlation is quite good at 85.9 % linearity (coefficient of determination). The regression of these two scales is used to derive a fitted (calibrated) free energy of transfer and reported in Table 1 and used in our new alignment algorithm. Since Gtow reflects a delta G (energy) of transfer, hydrophobic proclivities can also be seen to relate directly to energy (Table 1).

**Hydrophobicity and protein folded structure**

The reasonableness of our hydrophobicity scale is also demonstrated by examining the relationship between our scale and the mean residue depth (dpx) defined as the distance between the interior of a protein amino-acid and the nearest water molecule in the aqueous shell surrounding the protein (35,28). In Figure 3 we show that there is a strong relationship \((97 \% \text{ linearity})\) between the dpx metric and our hydrophobic proclivities. The dpx metric is a straightforward geometrical description of the local protein interior and can be expected to provide similar information to the solvent accessible area and buried surface area metrics. The dpx depth and hydrophobic proclivities correlate with amino-acid/protein properties such as average protein domain size, secondary structure, protein stability, free energy of formation of protein complexes, major literature amino-acid hydrophobicity scales, residue conservation, post-translational modifications like phosphorylation, and hydrogen/deuterium amide proton exchange rates (35,25,28).

**Derivation of property class vectors from scatter-plots**

Consider Figure 4 where our normalized average hydrophobicity index is scatter plotted against the area per specific volume of each amino acid (shown using their alphabetical representations).

We can clearly see cross-hatched patterns where for example the amino acids G, A, C, V, I and L are on a straight line (starting from the top left to bottom right). Moving right, we see that S, P, T, M and F are on a straight line (nearly parallel to the line formed by G,A,C,V and I). Continuing further right, we see a third line which crosses several amino acids, followed by an outlier, amino acid R. This series of four lines form what we call **Property Class 1**. We assign a numerical value of 0 to the line through G,A,C,V and 1 and a value of 1 to the next line and so on. In the same Figure 4 we can see the formation of **Property Class 2** which contains only two linear clusters ranging from the lower left to the upper right. One of the Class two series consists of V, I, L, C, M, P, Y and W, with the other series made up of the rest of the amino acids. We arrived at **Property Class 3** and **Property Class 4** by scatter plotting our normalized average hydrophobicity index against specific absolute entropy (and this is shown in Figure 5). The four property classes we identified respectively in the scatter plots shown as Figure 4 and Figure 5, along with the respective X axes physico-chemical property, correlated very highly (as multiple linear regression factors) with our normalized average of three robust hydrophobic indices (shown as avg 3H) having an R squared >95 %. Property class #5 reflects a scatter plot between the delta
Table 1. Table of Regression Fitted Hydrophobic Proclivities.

| Residue Amino Acid | Hydrophobicity (H) | Regression Fitted $\Delta G$ |
|-------------------|--------------------|-----------------------------|
| F (Phenylalanine) | 0.0688             | 2.5658                      |
| L (Leucine)       | 0.0579             | 2.6095                      |
| I (Isoleucine)    | 0.0349             | 2.7022                      |
| M (Methionine)    | 0.2213             | 1.9528                      |
| V (Valine)        | 0.1427             | 2.2687                      |
| P (Proline)       | 0.7123             | -0.0212                     |
| T (Threonine)     | 0.6599             | 0.1895                      |
| S (Serine)        | 0.7074             | -0.0018                     |
| A (Alanine)       | 0.4925             | 0.8624                      |
| Y (Tyrosine)      | 0.4523             | 1.0237                      |
| H (Histidine)     | 0.6763             | 0.1232                      |
| Q (Glutamine)     | 0.8692             | -0.6522                     |
| N (Asparagine)    | 0.8350             | -0.5148                     |
| K (Lysine)        | 0.9651             | -1.0376                     |
| D (Aspartic Acid) | 0.9157             | -0.8393                     |
| E (Glutamic Acid) | 0.8974             | -0.7657                     |
| C (Cysteine)      | 0.2650             | 1.7769                      |
| W (Tryptophan)    | 0.3403             | 1.4742                      |
| R (Arginine)      | 0.9091             | -0.8126                     |
| G (Glycine)       | 0.6582             | 0.1961                      |

Table 2. Linear correlation between hydrophobicity scales and AA physico-chemical properties.

| H Scale         | Moelbert ASA $R^2$ | F & P $C_{\text{OH}} R^2$ | 8 AA Property Class $R^2$ | 6 factor $R^2$ |
|-----------------|--------------------|---------------------------|---------------------------|---------------|
| Chothia         | 44.1%              | 58.6%                     | 86.0%                     | 88.6%         |
| Kyte-Doolittle  | 61.9%              | 65.7%                     | 97.6%                     | 94.8%         |
| Jannin          | 56.2%              | 68.3%                     | 82.7%                     | 80.7%         |
| Juretic Avg     | 63.7%              | 69.2%                     | 98.0%                     | 94.9%         |
| SEA >30         | 70.7%              | 72.7%                     | 90.4%                     | 89.4%         |
| Engleman-Steitz | 53.0%              | 72.8%                     | 79.4%                     | 87.4%         |
| Eisenberg-Weiss | 56.4%              | 76.1%                     | 87.8%                     | 71.3%         |
| Rose Avg % buried | 86.1%         | 81.7%                     | 85.6%                     | 86.8%         |
| Hopp-Woods      | 71.7%              | 82.7%                     | 66.5%                     | 71.3%         |
| Tang Q          | 86.0%              | 84.6%                     | 97.4%                     | 91.3%         |
| avg 3H          | 84.7%              | 85.9%                     | 98.23%                    | 95.8%         |
| Neumeirer X     | 90.2%              | 89.1%                     | 96.7%                     | 94.2%         |
| F & P del G $C_{\text{OH}}$ | 85.3% | 100.0%                   | 89.4%                     | 89.6%         |
The threshold of statistical significance for the coefficient of determination is 68 percent.

In Table 2 we summarize the performance of several of the hydrophobicity scales published in the literature. The hydrophobicity scales shown as rows are compared with four important quality metrics that are either amino acid physico-chemical properties or derived from such properties. The quality of inter scale regressions are shown as $R^2$. The performance of each row scale can be observed relative to the other row scales within each of the four columns, where the higher the $R^2$ the better the performance of the row scale with regards to the column scale. There are 13 rows in Table 2 representing 11 hydrophobicity scales, one solvent exposed area scale and one delta G of transfer from water to an organic solvent (Octanol).

### MLR regression $R^2$ assessments of hydrophobicity scales and proxies thereof

Of the 11 hydrophobicity scales in Table 2, 7 are popular scales in practice, three are the constituent scales of our hydrophobicity proclivity scale and our hydrophobicity proclivity scale. These row choices in Table 2 are to illustrate a close relationship between AA hydrophobicity and the transfer of an amino acid to an organic solvent (n-Octanol, Table 1 column 2), used as a proxy for the internal environment of a folded protein, as well as to compare AA hydrophobicity with an AA Solvent Exposed Area scale (column 1) also representing a folded protein environment. The high $R^2$ between the row dG of transfer to Octanol and the first column AA Solvent Exposed Area (SEA) scale in Table 2 illustrates the aptness of comparing the dG of AA burial in protein “solvent” to a solvent-solvent transfer model between water as the reference state and an organic solvent as the transferred or final state. In Table 2, the inclusion of the row SEA is to illustrate the high $R^2$ with the first column SEA illustrating the consistency of folded protein behaviour in SEA scales derived from different data sets. With the Rose AA percent buried row hydrophobicity scale, similar lessons can be gleaned as with the row Octanol and SEA scales, as the Rose scale represents the environment of a folded protein. The very high $R^2$ between these three row scales and the last two column regression scales in Table 2 illustrate a strong justification for including these row scales, as protein folding is thereby strongly linked with other physicochemical properties of amino-acids, as reflected by these two columns. We describe the regression X variables in the 4 columns of Table 2 below.

We can see that the correlation between our hydrophobicity scale (shown as avg 3H in Table 2) and the Moelbert average amino-acid solvent Accessible Surface Area (ASA) within
Figure 3. Hydrophobic Proclivities versus Structure based mean residue depth (dpx). The Y scatter about the regression line is pretty low yielding a regression $R^2$ of 97%.

Figure 4. Hydrophobic Proclivities versus Area per specific volume of amino acids. There are two Property Classes illustrated here with 4 red and 2 blue lines with each line assigned an ordinal number relative to zero within its property class.
proteins has an $R^2 = 84.7\%$. The ASA is the average area of each amino acid exposed to water in the globular proteins. When our hydrophobicity proclivity scale approaches 1 (i.e. hydrophilic) the ASA goes up as would be expected, with the converse being true as our hydrophobicity scale approaches 0 (i.e. hydrophobic) the ASA goes down\(^2\).

The amino-acid Accessible Surface Area (ASA) has long been suggested as a reasonably accurate proxy for hydrophobicity\(^{15,22,27}\) as is also seen in a related scale, the Solvent Exposed Area > 30 square angstroms\(^{22,23}\). The amino-acid property classes are vector sets of clusters/linear families of curves in multiple linear regression relationships between two (or more) amino-acid physico-chemical properties. The first two columns (ASA and Gtow) represent paired variable linear regressions and the third column (Property Classes #1 to #8, Table 3) and fourth column (Property Class #1 to #4, AA area/specific volume\(^{10}\) and specific absolute entropy\(^{8,9}\)) represent multiple linear regressions.

The $R^2$ in the first two columns of Table 2 represent linear regression results between the Y (row) vectors and the X (column) vectors. The $R^2$ in the last two columns of Table 2 derive from multiple linear regressions, where the independent (X) variables are vectors of amino-acid property Classes (PC) and/or amino-acid physico-chemical properties, and each row parameter is the dependent variable, respectively. Again, the Property Classes can be thought of as distinct subsets of amino-acids representing multiple linear series/clusters (within scatter plots or multiple linear regressions) of amino-acids in reference regressions associated with X variable vectors from some key physico-chemical metrics plotted against the hydrophobicity proclivity vector scale.

In Table 2, we see that the F and P Gtow scale performs as well (i.e. high $R^2$) as the best of the hydrophobicity scales within columns 1, 2 and 4, thus, further justifying our selection of the Gtow scale as our baseline standard for a free energy of transfer from an aqueous solvent environment to a non-aqueous solvent. The SEA > 30 $A^2$ does as well as the popular hydrophobicity scales in Table 2 and has good correlation with the F and P Gtow scale in column two and thereby establishes a direct link between the F and P Gtow scale and the free energy of burial of amino-acids in proteins and providing strong evidence justifying a solvent-solvent transfer model for protein folding.

The Tang Q and Neumeier X scales are the top performing individual hydrophobicity scales as seen in the first two column results, followed on average by the Rose scale. The Juretic Avg scale generally performs as well as the five popular hydrophobicity scales in columns one and two, but more importantly it performs better than any other single hydrophobicity scale except for the Tang Q and Neumeier X scales in columns three and four. Since we consider columns three and four to be a more rigorous test for a robust, high performance hydrophobicity scale,

Figure 5. Hydrophobic Proclivities versus specific absolute entropy. There are two Property Classes illustrated here with 4 blue and 3 red lines with each line assigned an ordinal number relative to zero within its property class.
we see the justification for selecting the Tang Q, Neumeier X and Juretic Avg as the scales from which to prepare our hydrophobicity proclivity (3H) scale. Our hydrophobicity proclivity scale performs basically as well as the best individual hydrophobicity scales in columns one and two, but it is the top performer in columns three and four. No other hydrophobicity scale that we evaluated on average performed as well (i.e. magnitude of R²) in regression comparisons with amino-acid physico-chemical properties as our hydrophobicity proclivity scale. In Table 2 column three is the 8 sets of numbers (vectors), dubbed as property classes and are eight X vectors (Table 3) in the multiple linear regression relationships with the R² shown in the third column. These eight property class vectors can form multi-linear regression fits with very high R² with a large number of the physico-chemical properties of the of the 20 amino-acids in our accumulated AA physico-chemical property database, thereby serving as proxy’s for these properties. In Table 2 column four, we see four property class vectors (#1-#4) and two AA physicochemical property vector scales (surface area/specific volume, specific absolute entropy); column four is included to illustrate the method of construction of the eight Property Class (PC) vectors represented by column 3.

Wolfenden and Carter dG of transfer from vapor to solvent, and water to cyclohexane of AA secondary group analogs

Strong relationships exist between some of our aminoacid property database and the Wolfenden free energy (delta G) of transfers between vapor-water, vapor-cyclohexane and cyclohexane-water. This experimental system using dG of transfers of amino-acid secondary group analogs was devised to eliminate the confounding effect of water content within the organic solvent and to eliminate the effect of the amino-acid backbone, such that only the contributions of the secondary group analog dG free energies of transfer could be calculated. The dG of transfer of vapor-water is strongly correlated to several hydrophobicity scales, drawing a link with the rest of the cognate literature, and the amount of polar area in the respective amino-acids (Table 10). This latter correlation is an important insight into the energetic distinction between polar/non-polar surface area as part of the hydrophobicity phenomenon. The dG of AA secondary group analog transfer between cyclohexane-water also strongly correlates to several hydrophobicity scales, but also correlates to average amino-acid exposed (to water) surface area in folded proteins (Table 10). Also notice the strong relationship between dG of transfer of AA secondary groups between vapor-water and cyclohexane—water in scatter-plot Figure 17. The dG of AA secondary group analog transfer between vapor and cyclohexane is strongly related to amino-acid property scales related to the size, bulk, mass and polarizability of the amino-acid secondary groups (Table 10).

The data discussed above is from our database and joint papers by Dr Richard Wolfenden, Dr Charles Carter and/or Dr. Anna Radzicka. The joint papers omit the data for Proline. The dG of transfer approach using classical physical chemistry procedures relating the vapor phase, Cyclohexane and water is unique in the literature. The data estimates for Proline are based upon multiple AA property regressions, because the data is not covered in original work for several reasons. First off, the secondary group of 3 methylene Carbons forms a 5 membered ring between the alpha Carbon and the amino group attached to the alpha Carbon. Secondarily, Proline is not technically an amino-acid, so the structural difference impacts its interaction energy with water and Cyclohexane. Thirdly, the actual preparation of a Proline analog is not without difficulty. Several estimates of key Proline values were undertaken through regression modeling. Without Proline in the data, the regression R² estimates don’t change very much from the regressions with and without the estimated Proline data values.

1D ANOPA analysis

Plot Figure 6 is the A0 distances histogram, which represents the distances from the centroid average to each pattern vector/point. This 1D ANOPA analysis is independent of any input to the calculations and is used as the first step in the algorithmic selection of the best pair of pattern points (objects) to select for the out group average for the 2D and 3D ANOPA analysis. We see four putative sub-populations (perhaps non-random correlation sub-structures) representing correlation based clustering. The ANOPA algorithm seeks to find pairs of pattern points from either the right hand side sub-population distribution tail (first pass) or the left hand side sub-population distribution tail (second pass, if needed) that represent an out group choice that is generally congruent in number and size of the 1D ANOPA histogram sub-populations, which in this case is 4, with the 2D ANOPA sub-population clusters in Figure 7. The green line represents the selection of the amino-acids Leucine and Isoleucine on the right tail of the 3rd sub-population. The Sub-populations are processed from right to left in the iterative search for the best two amino-acids to form the out group average.

2D ANOPA analysis

With (I, L) chosen for the out group, the 2D ANOPA (T0, d2) scatter plot Figure 7 reveals 4 clusters (defined by oval circumscribing) and a singleton point. This pattern in the 2D ANOPA scatter plot is consistent with the number and size of sub-populations in the 1D ANOPA histogram analyses. T0 are the normalized parametric distances along the relation vector/line, where the out group average is normalized to 1. The d2 distances are the pattern point distances from the relation vector/line. The clusters in the 2D ANOPA scatter plot Figure 7 are resolving by hydrophobicity, polarity, charge, degree of secondary group SP² hybridization and secondary group size. The 3D ANOPA plot is in Figure 11.

3D ANOPA analysis - Z’ vs. X’

Each distinct pair of the 3D ANOPA coordinates forms one of the 3 coordinate planes in Plot Figure 8. These 3 pairs of coordinates for each amino-acid effectively provides a projection of the 3D structure onto each of the three coordinate planes. In this plot there are 4 distinct clusters of amino-acids numbered 0-3. Scatter plots and correlation analysis between Z’ and a number of hydrophobicity scales show a very strong linear/curvilinear correlation, which is not surprising given that many of the 49 AA property scales are either hydrophobicity scales or
Figure 6. 1D ANOPA histogram. There are 4 population clusters and a single value as shown and this structure is consistent with the structure of the 2D ANOPA sub-population analysis.

Figure 7. 2D ANOPA plot. T0 vs. D2 distances. There are 4 population clusters and a single point as shown and this structure is consistent with the structure of the 1D ANOPA A0 distance histogram sub-population analysis.
strongly correlate with hydrophobicity scales, like reversed phase HPLC retention times. Reverse phase HPLC retention times represent the interaction between an AA and a water/non-polar surface where a clathrate structure forms. Correlation analysis with the $y'$ coordinates with their corresponding $d^2$ scales are very linear ($R^2 = 97.27\%$); the later representing the amino-acid distances from the relation vector. The $Y'$ coordinate represents the relative amount of polar area and aliphatic vs. aromatic/SP$^2$ hybridization area. See Table 6 and Table 4.

3D ANOPA analysis - $X'$ vs. $Y'$
Within each of the 4 clusters that we see in Plot Figure 9, the most polar/charged secondary group is located toward the top with the progression moving to the lesser polar/degree of charge, possessing more hydrophobic area, on the lower end. The AA size and amount of hydrophobic surface area is a large part of what drives these clusters, with there being a difference in the hydrophobicity of Carbons in different hybridization states with a SP$^0$ (aliphatic) hybridized Carbon more hydrophobic than a Carbon in a SP$^2$ (aromatic) hybridization state. Sulfur atoms act more like hydrophobic area in amino-acids compared to what can be seen in band 2 above with Nitrogen or Oxygen atoms, although not as strongly hydrophobic as Carbons in a SP$^3$ configuration. The hydrophobicity of an amino-acid is in direct relation to the total amount of hydrophobic surface area and the relative area partitioning between aliphatic Carbon, aromatic Carbon and Sulfur. The $X'$ coordinates also have a reasonably strong linear relationship to secondary structure revealed in scatter plots against metrics sensitive to secondary structure, such as a metric derived from our hydrophobicity proclivity scale, a beta sheet configuration proclivity scale, and a double bend proclivity scale (2*H-B-DB), which has good performance compared to the secondary structure of proteins of known secondary structure (unpublished results). When a plot of the primary sequence ordinal numbers against a running average of size 3 of the 2*H-B-DB metric goes above 0 (especially above 0.5) an alpha helix is indicated and when the running average of 3 of this metric goes below zero (especially below -0.15) a beta sheet is indicated. This metric is symmetric about zero and has a nice sinusoidal shape in regions of intercalating and alternating alpha helices and beta sheets or intercalating and alternating alpha helices. See Table 4 and Table 5.

3D ANOPA analysis - $Y'$ vs. $Z'$
The $Z'$, $Y'$ Plot Figure 10 has the hydrophobicity related $Z'$ ordinate as the X-axis and the $Y'$ ordinate picks up something of a non-linear/curvi-linear component involving the relative amounts of polar surface area, with the relative fraction of dipole vs. ionic surface area, with non-polar surface area and with the aliphatic vs. aromatic area fractions. We can see the former effect in the main sequence of group #1, in the relative placement of group #0/group #4 and in the placement of Tyrosine above Phenylalanine in group 4. See Table 5 and Table 6.

3D scatter plots. Two classes of amino-acids
In the three 3D ANOPA Plots (Figure 12 and Figure 13 and Figure 14) we see the relation vector, ensemble average to the out-group average L,I, and its reflection, with a total line segment.
**Table 4.** *X'* strong correlations. We see that the ANOPA *X* is reflective of protein secondary structure propensities, which interestingly have some bearing on HPLC retention times. The reflection of secondary structure by the *X* ordinate may be because the information reflected by the *X* ordinate is related to Van Der Waals dispersion forces and the secondary group induction effects that modulate the pk-C of the amino-acid Carboxilic acid groups, both effects of which can be related to the fraction of non-polar surface area and/or polar surface area.

| Property | Value |
|----------|-------|
| Retention time TFA - HPLC | beta-sheet |
| H2On free energy pk-C | Van Der Waals R0 |
| Van Der Waals epsilon | Retention time HPLC pH2.1 |
| Retention time HPLC NaClO4 | Retention time HPLC NaH2PO4 | HPLC parameter |
| RF high salt chromatography RF rank | Chothia hydrophobicity ANOPA Ar |
| Kidera Beta sheet avg Kidera double bend | Polar surface area =rose area-NP area |
| 49 property PCA Axis I | 49 property nmMDS Axis I |

**Table 5.** *Y'* strong correlations. The *Y* ANOPA ordinate is strongly related to a couple of HPLC scales and a couple of ANOPA process related distances. The *Y* ordinate has a relationship with PCA ordinate II and nmMDS ordinate II. Through these latter two relationships the ANOPA *Y* ordinate has a relationship with size and surface area, which are reflected as links with some of the hydrophobicity metrics.

| *Y* strong correlations | Value |
|------------------------|-------|
| ANOPA A0 ANOPA D2 | waters 02 AA+ACQ - HPLC, Waters 05 AA+ACQ - HPLC |

| *Y* correlations through nmMDS/PCA correlations | Value |
|-----------------------------------------------|-------|
| Optical rotation, Jannin hydrophobicity | Chothia hydrophobicity, Probability of Core, #methylenemethylene/methyl |
| 5 scale 'normalized' average hydrophobicity | prob (core) #46, avg % buried #89, Normalized consensus Tang Q scale hydrophobicity |
| Juretic avg hydrophobicity Hydrophobicity Proclivity - avg best 3H Miyazawa & Jurnigan Eir | Miyazawa & Jurnigan -#6*Eir*Qi |
| Fauchere & Pliska in Karplus del G C8OH kcal/mole Kidera PCA free residue H Scale 2 | Kidera PCA buried AA H Scale 3 |
| Rose In(%buried) & Miyazawa Eir (normalized consensus), Wimley F&P corrected dG transfer C8OH | 49 property PCA Axis II |
| 49 property nmMDS Axis II |

The *Z'* axis is primarily related to hydrophobicity. In the background in gray is the 0,0,0 origin marked by a small (X,Y,Z) coordinate axes. The 3D ANOPA procedure calculates the amino-acid pattern point projection distances along the relation vector, the pattern point distances from the relation vector and the angle of rotation about the relation vector. In recent work by Dr. Charles Carter and Dr. Richard Wolfenden\textsuperscript{28-33}, they have found that their hydrophobicity measures/scales partition into two parts reflecting the amino-acid assignments to the two types/classes of amino-acid tRNA Synthetase (aaRS) enzymes that charge the tRNA's with their matching amino-acids. The Carter/Wolfenden work shows that the anti-codon loop of tRNA's reflect the hydrophobicity dual partitioning as does their newly discovered and more primitive tRNA stem code. The amino-acids have been color coded green or purple depending upon which of the two aaRS's their matching tRNA's map to. We can see that the purple points cluster together and are sandwiched between the green points, where the purple points represent amino-acids of either moderate hydrophobicity/size and the smaller polar amino-acids Aspartic acid and Asparagine. These last two 3D ANOPA views give more perspective on the placement of amino-acids with respect to hydrophobicity and by whether or not they partition according to their class I or class II aaRS enzymes.
Table 7. ANOPA, PCA, nmMDS analysis. Statistically significant correlations are in bold face and weaker, non-trivial correlations in italics. Each pair of columns represent an ANOPA axis that is correlated with the respective PCA or nmMDS axes. Each pair of columns represent a linear correlation and a curvi-linear correlation, with the latter typically being higher. The correlations are expressed as $R^2$, which is more conservative statistic. With one exception, each ANOPA ordinate only corresponds with one PCA or nmMDS ordinate. The regressions from which the $R^2$ derive are from a simple linear regression (plain $R^2$) and an unconstrained sixth order polynomial curvi-linear regression ($O(6)$ Poly $R^2$).

| ANOPA | X’ | Y’ | Z’ |
|-------|----|----|----|
| nmMDS | Linear $R^2$ | Poly O(6) $R^2$ | Linear $R^2$ | Poly O(6) $R^2$ | Linear $R^2$ | Poly O(6) $R^2$ |
| Axis I | 82.2 | 88 | 0.3 | 33.6 | 0 | 31.3 |
| Axis II | 0.2 | 30.4 | 19.3 | 68.3 | 68 | 74.1 |
| Axis III | 0.1 | 48.6 | 0.6 | 46.3 | 58.6 | 71.2 |

| PCA | | | |
|-----|----|----|----|
| Axis I | 79.5 | 88.6 | 1 | 30.1 | 3.1 | 31.6 |
| Axis II | 5 | 53.2 | 3.9 | 50.2 | 82.2 | 90.6 |
| Axis III | 2.3 | 20.7 | 10.6 | 59 | 11.5 | 32.7 |

Figure 9. X’ vs. Y’ 3D ANOPA plot. This plot represents a property class with 4 clusters of amino-acids each assigned an ordinal number relative to zero.
**Meaning of ANOPA X', Y' and Z' axes**

What we see in the ANOPA (Z', Y'), (Z',X') and (X',Y') Plots (Figure 8, Figure 9 and Figure 10) are the relationship between each amino-acid and the cavity it creates in water. The creation of a cavity in water creates a clathrate shell of waters with a surface tension. There is a pressure-volume work energy cost for creating the cavity in water and a surface tension-area work energy cost in opening of the cavity in water. We also see that there is an offsetting free energy cost that mitigates the aqueous cavity opening cost that derives from the attraction (wetting force) of the aqueous clathrate shell to the surface of the amino-acid. We have non-polar surface area (aliphatic, aromatic/PI resonance network) interacting through Van Der Waals and induced dipole interactions. We also have dipole, hydrogen bonding, and ionic interactions between amino-acid surfaces with the aqueous clathrate shells. The surfaces of amino-acids are partitioned up to have some subset of these interactions with the aqueous clathrate shells. Table 4, Table 5 and Table 6 are the summary...
Figure 12. 3D ANOPA plot with L & I as the out-group. The data points are green or purple depending upon which aaRS class (I,II) color code they correspond to.

Figure 13. 3D ANOPA plot with L & I as the out-group. The data points are green or purple depending upon which aaRS class (I,II) color code they correspond to.

Figure 14. 3D ANOPA plot with L & I as the out-group. The data points are green or purple depending upon which aaRS class (I,II) color code they correspond to.
results of a correlation study of the ANOPA X’, Y’ and Z’ versus correlated/highly correlated scales from our amino-acid property database.

The correlation between the Z’ axis and hydrophobicity proclivity scale means that there is a strong justification for this H-index being a preferred hydrophobicity scale. Many of the amino-acid properties within the 49 AA property scales used for the ANOPA analysis are either hydrophobicity scales or are scales partially to strongly related to the factors underlying hydrophobicity.

We have used the rules of ANOPA to look at choices of amino-acids from the outer and inner sub-population tails in the 1D ANOPA A0 distance histogram plot Figure 6. The deliberate selection of a number of amino-acid properties related directly or indirectly strongly suggested that a pair of large surface area aliphatic residues would be an optimal choice as it was found through the ANOPA 2D and 3D analysis and other subsidiary analysis.

**Principal Components Analysis (PCA) and non-metric Multi-Dimensional Scaling analysis (nmMDS)**

We can cross check the ANOPA analysis through a Principal Component Analysis (PCA) analysis of the same 49 amino-acid property scales and by calculation of correlation coefficients between the 3 axes of each of the PCA analysis (3D scatter-plot Figure 15) and ANOPA analysis. To even add more grist for the mill to confirm the ANOPA findings we can also run a non-metric Multiple Dimensional Scaling (nmMDS) analysis (3D scatter-plot Figure 16) on the same 49 amino-acid property scales. We see the results in the two coefficient of determination ($R^2$) Table 7 and Table 8, where we find that there are good to excellent correlations between the three analysis. A strong correlation between multi-variate classification procedure ordinate scales are reflected in bold fonts and weaker, but non-trivial, relationships are reflected in non-bold, italic font. The inter-ordinate axis correlated relationships are measured by the coefficient of determination ($R^2$), both linear correlation and curvi-linear correlation with polynomials of order 6 – O(6). Generally speaking, there is good correlation amongst the three coordinate systems as seen in the ANOPA vs. nmMDS/PCA Table 7 and in the nmMDS vs. PCA Table 8. Where there is a strong linear correlation there is also a strong curvi-linear correlation, with the latter correlation scoring higher. In these two tables, with the exception of the ambiguous ANOPA Y’ scale relationships with the nmMDS Axis II and III scales, the rest of the relationships have singularly strong paired relationships, whereas the other paired relationships are weak, if not zero (independent). In plot (Figure 18) it can be seen that there is a fast roll off of the PCA Eigenvalues/variances, with the first three principal components representing 68% of the pattern point variation. ANOPA suffers no such loss of information.

From this perspective of the 3D nmMDS plot Figure 16, we see that Alanine (A) is the far apex past the ensemble average, from which the point cloud fans out in moving toward the viewer, somewhat akin to a cone. The points position themselves according to hydrophobicity/polarity, alky1 vs. aromatic surface area and size. The amino-acids L,I,V,C are located off to one side as in the 3D ANOPA analysis.

---

*Figure 15. 3D PCA plot.*
Table 8. nmMDA vs PCA ordinate correlation analysis. There is only one strong and statistically significant match between each of the 3 PCA and 3 nmMDS axes, which provides strong evidence to conclude that these two analysis are essentially telling the same story and reinforcing the comparison between these two multi-variate analytical methods and ANOPA. The regressions from which the $R^2$ derive are from a simple linear regression (plain $R^2$) and an unconstrained sixth order polynomial curvi-linear regression ($O(6) R^2$).

| nmMDS vs. PCA | Axis I | Axis ii | Axis iii |
|---------------|--------|---------|----------|
|               | Linear $R^2$ | Poly O(6) $R^2$ | Linear $R^2$ | Poly O(6) $R^2$ | Linear $R^2$ | Poly O(6) $R^2$ |
| Axis I        | 96.01   | 97.88   | 3.28     | 34.44     | 0.05     | 12.97 |
| Axis ii       | 2.14    | 17.61   | 37.3     | 44.51     | 48.3     | 67.77 |
| Axis iii      | 6.18    | 30.08   | 83.58    | 88.06     | 9.11     | 46.75 |

Table 9. Property Class inter-scale correlations are shown below. Since the PC values are discrete, ordinal values that encode sub-set membership and the fact that due to the nature of the amino-acids there are going to be some sub-set memberships that overlap, we can expect to see some inter-scale correlation. An inter-scale correlation would have to exceed 0.806 to be statistically significant and there is only one pair of scales that meet this condition. For reasons discussed in the method section, the apparent correlation does not reflect a significant loss of dimensionality and there is a different informational content reflected by each scale. The MLR regressions will servo/weight each PC scale to effectively include/exclude the scale from being related/utilized in the establishment of a MLR regression correlation used to judge the quality of relationship of various amino-acid property/attribute scales.

| Pairwise correlation coefficient table |
|---------------------------------------|
| PC # 1 | PC # 2 | PC # 3 | PC # 4 | PC # 5 | PC # 6 | PC # 7 | PC # 8 |
| ----- | ------ | ------ | ------ | ------ | ------ | ------ | ------ |
| PC # 1 | 1      |        |        |        |        |        |        |
| PC # 2 | 0.396  | 1      |        |        |        |        |        |
| PC # 3 | -0.629 | -0.061 | 1      |        |        |        |        |
| PC # 4 | 0.444  | 0.860  | 0.150  | 1      |        |        |        |
| PC # 5 | -0.592 | -0.769 | 0.067  | -0.753 | 1      |        |        |
| PC # 6 | -0.488 | -0.411 | 0.244  | -0.554 | 0.433  | 1      |        |
| PC # 7 | 0.129  | 0.122  | -0.089 | 0.057  | -0.245 | 0.481  | 1      |
| PC # 8 | -0.580 | -0.549 | 0.285  | -0.651 | 0.667  | 0.696  | -0.089 | 1      |
Table 10. Database items strongly related to the Wolfenden-Carter dG vapor-liquid, liquid-liquid, amino-acid secondary group analog dG of transfers. These statistically significant relationships provide us with more insight into the meanings of both the Wolfenden-Carter dG scales and the database metrics reported in this table. The v>wa scale is the dG of vapor to water. The c>wb scale is the dG of Cyclohexane to water. The vap - chx scale is the dG of transfer of vapor to Cyclohexane. The transfer dG's of Cyclohexane to water and vapor to water are strongly correlated. The transfer dG of vapor to water is strongly related to hydrophobicity and polar surface area. The transfer dG of Cyclohexane to water is strongly related to surface area and proclivity to have more than 30 square angstroms in contact with water in folded proteins. The dG of transfer from vapor to Cyclohexane is strongly correlated to molecular size and asymmetry. The regressions from which the R² derive are from a simple linear regression (plain R²) and an unconstrained sixth order polynomial curvi-linear regression (O(6) Poly R²).

| Database # | Carter dG | Database # | AA Property         | Linear R² | Poly O(6) R² |
|------------|-----------|------------|---------------------|-----------|--------------|
| 235        | v>wa      | 36         | Steitz H scale      | 77.57     | 84.45        |
| 235        | v>wa      | 38         | Doolittle H Scale   | 73.78     | 88.11        |
| 235        | v>wa      | 39         | Jannin H scale      | 65.01     | 81.74        |
| 235        | v>wa      | 40         | Chothia H scale     | 76.78     | 84.10        |
| 235        | v>wa      | 41         | Weiss H scale       | 81.78     | 86.45        |
| 235        | v>wa      | 45         | Hydrophobe          | 73.32     | 84.77        |
| 235        | v>wa      | 167        | Polar surface area  | 87.20     | 93.73        |
| 236        | c>wb      | 36         | Steitz H scale      | 86.89     | 92.12        |
| 236        | c>wb      | 39         | Jannin H scale      | 70.43     | 90.27        |
| 236        | c>wb      | 41         | Weiss               | 93.76     | 98.05        |
| 236        | c>wb      | 47         | SEA > 30 Å²         | 71.08     | 90.23        |
| 236        | c>wb      | 87         | Avg avail surface area | 65.12  | 83.66        |
| 236        | c>wb      | 235        | v>wa                | 87.26     | 90.09        |
| 239        | vap - chx (d) | 43       | Daltons             | 92.97     | 94.33        |
| 239        | vap - chx (d) | 86       | area A²             | 84.80     | 89.70        |
| 239        | vap - chx (d) | 91       | area/spec vol       | 90.28     | 94.01        |
| 239        | vap - chx (d) | 124      | polarizability      | 87.61     | 91.58        |
| 239        | vap - chx (d) | 127      | Lc                   | 83.94     | 89.62        |
| 239        | vap - chx (d) | 128      | AA size             | 87.69     | 90.30        |
| 239        | vap - chx (d) | 138      | size/bulk           | 87.13     | 91.20        |
| 239        | vap - chx (d) | 172      | volume              | 76.76     | 82.65        |
| 239        | vap - chx (d) | 173      | surface area        | 87.79     | 90.56        |
| 239        | vap - chx (d) | 179      | Refractivity        | 70.44     | 88.46        |
| 239        | vap - chx (d) | 181      | centroid side chain | 84.93     | 88.32        |

Generally, we see the same clusters of amino-acids in the PCA 3D plot as we do in the nmMDS plot, such as K,R,D, V,I, F,W and P,T,Y. Broadly speaking, the amino-acid points position themselves according to hydrophobicity/polarity, alkyl vs. aromatic surface area and size similar to the nmMDS analysis. The amino-acids L,I,V,C are located off to one side as in the 3D ANOPA analysis.

Discussion
Hydrophobicity, clathrate condensation and a hypothetical mechanism associated with protein folding
We point out the excellent linear relationship between our hydrophobicity proclivity scale and the dpx average residue depth (from water) scale in Figure 5 of the manuscript. The dpx scale
Table 11. Property Class 1 to 8 MLR correlations from the third column of Table 2, with showing the F and T test statistical significance. A few relevant rows have been added for additional correlation purposes. Generally speaking, the T test is more conservative than the Excel 2010 regression package F test.

| Ref # | Amino-Acid Scale | $R^2$ | $F$ test | $T$ test |
|-------|------------------|-------|----------|----------|
| 37    | Hydrophobicity Hopp - Woods | 66.52% | 6.258240% | 4.618453% |
| 36    | Hydrophobicity Engleman - Steitz | 79.40% | 0.656718% | 2.122334% |
| 39    | Hydrophobicity Jannin | 82.71% | 0.276640% | 1.612981% |
| 89    | Rose avg %buried | 85.58% | 0.110798% | 1.216479% |
| 48    | SEA < 10 Å² | 85.92% | 0.098053% | 1.172110% |
| 41    | Hydrophobicity Eisenberg - Weiss | 87.81% | 0.046932% | 0.938934% |
| 236   | Dr. Wolfenden, Proxy for polarity Cyclohexane to water, chx - wat (a) | 87.94% | 0.044289% | 0.922828% |
| 40    | Hydrophobicity dG 95% buried, Chothia | 88.01% | 0.043104% | 0.915392% |
| 120   | Fauchere & Pliska delta G H2O & C8OH (Karplus) | 89.44% | 0.022324% | 0.753199% |
| 47    | SEA > 30 Å² | 90.42% | 0.013471% | 0.648655% |
| 201   | Moelbert accessible surface area Moelbert ASA | 90.72% | 0.011345% | 0.617641% |
| 235   | Dr. Wolfenden, Proxy for area, Vapor to cyclohexane, vap - chx (b) | 91.94% | 0.054070% | 0.498098% |
| 86    | Rose area Å² | 95.06% | 0.000396% | 0.236311% |
| 88    | Rose avg buried | 95.78% | 0.000171% | 0.186498% |
| 76    | Neumaier X | 96.70% | 0.000045% | 0.128374% |
| 101   | 5 Normalized Hydrophobicity scales, average, rank order (#38-39, 41) | 96.85% | 0.000035% | 0.119541% |
| 221   | Area/spec volume A/(V/M) | 97.30% | 0.000015% | 0.094725% |
| 68    | Tang Q scale | 97.44% | 0.000011% | 0.087459% |
| 38    | Hydrophobicity Kyte - Doolittle | 97.63% | 0.000007% | 0.077955% |
| 94    | Juretic avg | 97.96% | 0.000003% | 0.062335% |
| 95    | preferred 3 avg best H | 98.28% | 0.000001% | 0.048172% |

is derived from a suite of proteins with actual solved folded structures, hence dpx is in effect a protein structural description of hydrophobicity where the deeper on average that a residue is buried, the more hydrophobic it is. Contrast the dpx concept with the ideas of average solvent exposed area, average buried area, percent buried, etc. When the relationships in Figure 3–Figure 5 (with supporting literature) are taken together, the hydrophobicity potential driving protein folding is suggested to be the traditional solvent-solvent partitioning model of protein folding where amino-acids partition between aqueous exposure and burial within the hydrophobic core of folded proteins. We believe that this traditional model of the primary driving force for globular protein folding is apt, but needs some updates and modifications that we develop in detail within the forthcoming manuscript - Hydrophobicity revisited: A Molecular Story. We are suggesting in the present manuscript that aqueous clathrates form about the hydrophobic areas of amino-acids in an unfolded state (a structural feature) and hydrophobic surface area on folded proteins possess a surface tension like a gas bubble in water. We are also suggesting that the physics driving the initial stages of protein folding is the same physics as the coalescing/condensing of bubbles of non-polar gaseous species or of fine hydrophobic liquid droplet dispersion’s in water. We expand upon this hypothesis in the forthcoming manuscript.

The great organizing principle embodied within the hydrophobicity proclivities (and implied by dpx), is that of a neo solvent-solvent partitioning effect, where the energetics of the solvent shell waters are the dominant effect in the energy balance. As with clathrates (ordered aqueous shells), which form spontaneously with hydrophobic molecules, there is a solvent shell of ordered waters that form spontaneously around solvated globular proteins. However, there is a confounding factor in trying to obtain an accurate hydrophobicity proclivity in that even
Figure 17. Wolfenden-Carter amino-acid secondary group analog. $X = dG$ vapor to water (#235) vs. $Y = dG$ Cyclohexane to water (#236) plot.

Figure 18. Principal Components Analysis scree plot. The first 3 Principal Components represent about 67% of the data variation.
the most hydrophobic protein will have some average solvent exposed area, so it is reasonable to postulate that there is some functional reason for exposure of some grease to the solvent. The presence of hydrophobic surface area causes an aqueous clathrate shell to form at that point perhaps effectively becoming part of the folded structure of the folded protein, possibly as a retaining structural element operating through surface tension and putting the interior of the globular protein under pressure. The importance of amino-acid hydrophobicity to the structure and function of globular proteins is critical to the function and survival of cells, a reality that is even reflected in the very structure of the standard genetic code.

The amino-acid codons are arranged/coded in such a way as to reflect the underlying hydrophobicity of the respective amino-acids. A careful analysis reveals that the genetic code has a built in redundancy through amino-acid hydrophobicity (in addition to codon redundancy) such that point mutations in a codon that yield a different codon tend to result in an amino-acid with similar hydrophobicity. It has been shown that the underlying amino-acid codon structure has a direct relationship with high quality hydrophobicity scales that are published in the literature.15

Our aim in developing a single hydrophobicity scale for the purposes of protein alignments was to try to develop a scale which reliably represented the central (average) hydrophobic tendency as a robust first order effect that allows simple, but meaningful paired comparison of aminoacids for homology relationships. Adding additional variables/properties would in our opinion detract from our goals of simplicity of calculation and utilization of what we believe is the primary first order effects/mechanics driving the initial stages of protein folding.

The derivative variables concept used in this paper refers primarily to ratios of molecular properties or directly measurable bulk properties of amino-acids. In this way, we derive normalized intrinsic thermodynamical properties versus extrinsic/bulk properties. For example, we find that the ratio of amino-acid surface areas to their volume has a strong relationship to hydrophobicity. Other derivative variables that we have used could be exemplified by estimating the water cavity volume/surface area using amino-acid volume and surface areas and some assumptions about the geometry of the aqueous cavity.

**Hydrophobicity and protein alignments.**

A legitimate question about the hydrophobic prclivity scale we have described is why our scale is superior to alignment score matrices such as PAM (Point Accepted Mutation)16, BLOSUM (BLOCK Substitution Matrix)17 or Gonnet18 that continue to be used for multiple protein alignments and database search alignments.

There are indeed several practical and theoretical problems with the use of these log odds score matrices for the alignment of divergent protein sequences. For example, BLAST and several of the major multi-sequence alignment programs like Clustal W use particular BLOSUM matrices as the default. BLAST uses BLOSUM62 as the default. Quotes from select papers have been summarized below to more clearly illustrate these problems.

The substitution matrices used by the alignment programs are generally log of Bayesian probabilities for two aminoacids I and J of the form:

\[
Q_{ij} = \frac{\text{prob}(I \rightarrow J)}{\text{prob}(I) \times \text{prob}(J)}
\]

The probability of occurrence of the 20 primary aminoacids is not the same throughout the domain/kingdoms of life, so this mathematical formulation can cause issues for identifying and aligning homologous proteins from distantly related organisms.

Superimposed on the log of Bayesian probabilities formalism are evolutionary models derived from Markov stochastic process evolutionary models (PAM), which implies _apriori_ knowledge of the evolutionary amino-acid substitution rates. Necessarily, if one chooses PAM or BLOSUM, one must choose one of the series of matrices that one believes is appropriate for the approximate evolutionary distance between any two protein sequences under analysis. Obviously, this practice can cause an undue restriction if the evolutionary distance is too great within the protein dataset being aligned. The only assumption that we make with hydrophobicity and our new alignment algorithm is that nature will strongly tend to substitute similar amino-acids in order to preserve the overall function and structure of homologous proteins, and that it is possible to define a hydrophobicity distance to define a fuzzy match between any two amino-acids, which is recognized as a "similarity match."

We summarize the salient points regarding alignment matrices with quotes from four select literature articles below.

1. “Many phylogenetic inference methods are based on Markov models of sequence evolution. These are usually expressed in terms of a matrix (Q) of instantaneous rates of change but some models of amino acid replacement, most notably the PAM model of Dayhoff and colleagues, were originally published only in terms of time-dependent probability matrices (P(t)). Previously published methods for deriving Q have used eigen-decomposition of an approximation to P(t). We show that the commonly used value of t is too large to ensure convergence of the estimates of elements of Q. We describe two simpler alternative methods for deriving Q from information such as that published by Dayhoff and colleagues.”

2. “The most common substitution matrices currently used (BLOSUM and PAM) are based on protein sequences with average amino acid distributions, thus they do not represent a fully accurate substitution model for proteins characterized by a biased amino acid composition.”

3. “We have investigated patterns of amino acid substitution among homologous sequences from the three Domains of life and our results show that no single amino acid matrix is optimal for any of the datasets.”

4. “Many phylogenetic inference methods are based on Markov models of sequence evolution. These are usually expressed in terms of a matrix (Q) of instantaneous rates of change but some models of amino acid replacement, most notably the PAM model of Dayhoff and colleagues, were originally published only in terms of time-dependent probability matrices (P(t)). Previously published methods for deriving Q have used eigen-decomposition of an approximation to P(t). We show that the commonly used value of t is too large to ensure convergence of the estimates of elements of Q. We describe two simpler alternative methods for deriving Q from information such as that published by Dayhoff and colleagues.”
4. These authors note another interesting problem with the residue substitutions rates use in the Q matrix: “Because different local regions such as binding surfaces and the protein interior core experience different selection pressures due to functional or stability constraints, we use our method to estimate the substitution rates of local regions. Our results show that the substitution rates are very different for residues in the buried core and residues on the solvent-exposed surfaces”.

Tomii et al. essentially conclude that in the “evolutionary” limit, alignment/mutation matrices reflect the hydrophobicity and amino-acid secondary group size. For example, when the correlation coefficient between a hydrophobicity scale and a amino-acid secondary group size, and the PAM matrices are plotted against the PAM distance, the correlation coefficient monotonically increases from 0.58 at a PAM near zero, to a PAM distance of 200 where the correlation coefficient reaches an asymptotic limit of about 0.73.

Conclusion
The amount of information available to an alignment algorithm is essential to its ability to find matching proteins, especially matches with remote homologies where the percentage identity has dropped off to around 20–25%. In this study we have sought to find an optimal, central tendency hydrophobicity scale that would reflect the real properties of amino-acids within the context of folded proteins. We contend that hydrophobic proclivities transcend mere statistical trends and reflect the functional necessities of globular proteins by amino acid properties according to a solvent-solvent (water interior of a folded protein) partitioning model. Within this model the primary driving force is that of water-water attractions that exceed water-amino acid attractions. Hydrophobicity is not a force that repels amino acids from water, but rather that water molecules attract each other more. When hydrophobic amino acids are exposed to water, clathrate shells spontaneously form at those areas, creating an anchored aqueous patch of ordered water molecules with surface tension. Thus, the preferred hydrophobicity scale of hydrophobic proclivities as we have described here provides significant new information to alignment algorithms and in particular our TMATCH algorithm (described elsewhere), optimized to work with our hydrophobicity proclivity scale.

There are some issues in using water/non-polar solvent partitioning ratios as proxies for hydrophobicity, primarily regarding the standardization of water concentration within a non-polar solvent (miscible with water), particularly when that solvent is capable of hydrogen bonding. Uncontrolled or variable amounts of water dragged into the organic solvent phase by solutes in solvent/water solute partition experiments can cause significant variation within the calculated Gibbs free energies of transfer of the solute from one phase to the other. We also point out that solvent/water solute concentration ratios, used to calculate free energy of transfers from water to a non-polar phase, as a hydrophobicity measure relevant to protein folding, suffers from a systemic error if the organic solvent is incapable of hydrogen bonding, which is not the case within the hydrophobic “solvent like” core of folded globular proteins. Our hydrophobicity proclivity scale has no units, but rather is a normalized proclivity. We show in Figure 4 of the manuscript that our hydrophobicity proclivity scale is directly relatable to free energy of solvent-solvent transfer using a popular dGow (water to n-octanol) scale. We also point out that characterizing the amino-acid secondary groups alone does not treat the important contribution to the free energetic’s of folding by peptide bonds, such as is accounted for by some researchers by treating the 20 amino-acids as guests in the center of tri-peptides in octanol to water free energy studies.

There is a large profusion of hydrophobicity scales in the literature, which pose many difficulties in trying to numerically reflect the wide range of amino-acid behaviors, as well as the difficulties inherent in trying to define hydrophobicity as an experimentally measurable concept related to the driving forces of protein folding. We have relied upon other experimentally measured amino-acid properties to cross check our hydrophobicity proclivity scale using methods such as Multiple Linear Regression (Table 3). For example, we can draw a good linear correlation between certain hydrophobicity scales, such as our hydrophobicity proclivity scale, and amino-acid reverse phase (C18 column) HPLC retention times; to which we point out that other researchers have drawn similar conclusions. We believe that the difficulties of using water to non-polar solvent dG of transfer have been largely mitigated and obviated through the analytical procedures reported in this paper.

In the final analysis, we are showing that a robust, high performance hydrophobicity scale enshrines much more information that can be captured within substitutional rate matrices, particularly for remote homologous proteins were the assumption of constant aminoacid substitutional rates become invalid for various reasons. Our TMATCH algorithm is structured to leverage the extra information inherent in our hydrophobicity proclivity scale to reflect a global alignment capturing the secondary and super-secondary structure which lead to tertiary structure. The TMATCH algorithm transition weighting scheme automatically tends to capture areas of high local hydrophobic similarity, despite low having low percent identity, like would be captured in a dot plot algorithm.

Author contributions
DC arrived at the hydrophobicity index several years ago after an exhaustive look at the literature and through extensive regression analysis of several published values of amino acid properties in proteins and how they may contribute to the structure of proteins in solution. This paper is the result of a long collaborative effort with KC whose interests were also in understanding protein structure and search algorithms in bioinformatics.

Acknowledgements
We (DC and KKC) appreciate our discussions about proteins and their solution structure with Dr. Joseph Ng (Department of Biological Sciences) and Dr. John Shriver (Department of Chemistry) of University of Alabama in Huntsville.
Ana Jerončić
Department of Research in Biomedicine and Health, University of Split School of Medicine, Split, Croatia

The authors aimed to develop a hydophobicity scale which optimally reflects properties of amino acids residues/amino acids that are relevant for folded proteins. The long term goal was to use such a scale to estimate hydrophobicity-based protein sequence relatedness from a global alignment in order to improve identification of structural homologs with less than 25% sequence identity.

The authors have identified eight properties (so called property classes) of amino acids relevant for folded proteins by recognizing distinct patterns in a series of scatter-plots that plotted many hydrophobicity indices and other physico-chemical properties of amino acids/amino acid residues against each other. Apparently, different property classes were visually recognized in scatter-plots after the “linear cluster of amino acids”-fingerprint was found (clusters of amino acids with similar physico-chemical properties that are aligned along distinct imaginary lines in a scatter-plot).

The final set of plots from which eight property classes were derived included: scatter-plots of the hydophobicity scale that was developed by the authors versus a) the area per specific volume of each amino acid (property classes 1 and 2) or b) the specific absolute entropy (classes 3 and 4); c) the plot of delta G of burial of amino acid secondary group versus number of atoms in a group (class 5); and finally, ambiguously defined “classes #6, #7 and #8 were derived from 49 fundamental aminoacid properties and derived scales that are based upon an analysis with Analysis of Patterns (ANOPA)".

It is this set of eight property classes that the authors have used as dependent variables in multiple linear regression models (MLR) of candidate hydrophobic scales. The measure of goodness of fit of a MLR model (R² value) was used to identify optimal hydrophobic scale(s) as according to the authors the MLR's R² value represents “rigorous test for a robust, high performance hydrophobicity scale”. The rationale for such assumption was that for all tested hydrophobicity scales, R² value of MLR models were higher than R² values of simple regression models using either Moelbert's average aminoacid solvent Accessible Surface Area (ASA) or Fauchere & Pliska
free energy of amino-acid transfer from n-Octanol to water (Gtow) as the dependent variable.

However, I disagree with the authors on the MLR $R^2$ rationale as I have concerns about appropriateness of data analysis (see below). In addition, the reporting in the manuscript should be substantially improved.

**Major comments**

**Data analysis**

**Comment 1**

Throughout the paper description of MLR models is very confusing and it is not clear what models were actually run (what was the dependent and independent variables; also the estimated coefficients and statistical significance of independent variables were not shown for a single model). As already stated, it seems that MLR models were mainly used to model different hydrophobic scales using eight *property classes* as independent variables. However on the page 7 the authors state “*The amino-acid property classes are vector sets of clusters/linear families of curves in multiple linear regression relationships between two (or more) amino-acid physico-chemical properties.*” which is very confusing – it seems that in addition to eight property classes, a few additional independent variables (amino-acid physico-chemical properties) were also included in a model or it was the multivariate regression model that was used (if yes – what were dependent variables)?

If indeed the authors used MLR models as claimed, this means that for a model with 20 observations (20 amino acids) at least eight independent variables were included in a model. Moreover, as majority of these *property classes* were actually ordinal variables, the number of independent variables included in MLR model should have been even higher (due to introduction of dummy variables). Consequently the sample size of these models was far too small to estimate model's parameters precisely, and the reported $R^2$ was actually quite inflated (due to an overfitted model but also a large number of independent variables that the authors did not adjusted for when comparing simple linear regression models and MLR). In addition, there was a problem of multicollinearity between independent variables which additionally inflated MLR's $R^2$ (based on the Table 2 Kendall tau coefficient between i.e. PC2 and PC4 is 0.82, $P<0.001$). Therefore the main result of this paper which is based on assumption that the property classes of amino acid residues identified through ‘linear-clusters' represent “real properties of amino-acids within context of folded proteins” is not based on validated assumption.

**Comment 2**

The authors have used a measure of goodness of fit of a regression model, $R^2$ to compare strength of linear relationship(s) modelled in different regression models (including simple and multiple linear regression models). However, $R^2$ is an overused statistics for linear regression analysis and additional metrics are required to get the whole picture. In particular, it is a Pearson correlation coefficient between paired data (i.e. two hydrophobicity scales) that quantifies the degree to which two variables are related and is a proper statistical measure of the strength of a linear relationship. Linear regression models find the best-fit line that predicts dependent variable from independent variable(s) with $R^2$ actually representing squared Pearson correlation of the fitted values and the observed values.

**Reporting**

**Comment 3**
A reader should know precisely which scatter plots were screened for “linear-cluster” pattern. This means that the entire set of hydrophobic scales and other physicochemical properties of amino acid residues/amino acids that were collected from the literature and used for generation of these plots should be listed in the paper. Also, it should be specified how many scatter plots were finally generated (in example: N*(N-1)/2 where N – number of hydrophobic scales or physicochemical amino acid properties that were collected, ...).

Comment 4
Since the eight property classes of amino acid residues are the most important novelty of this paper, the process of their identification should be clearly described in a sufficient detail. In particular:

1. What was the reasoning behind the assumption that the property classes of amino acid residues identified through ‘linear-clusters’ represent “real properties of amino-acids within context of folded proteins”. Or there was no assumption and the fact that the regression models of all hydrophobicity scales exhibited the highest R2 values when these property classes were used as independent variables actually justified such interpretation. If latter was the case, such reasoning would not be justified (see comments on multiple linear regression analysis)
2. The authors should describe the method they used to identify linear clusters on a plot (i.e. visual identification, followed by analysis of amino acid physicochemical/biochemical properties in clusters and regression-analysis of clusters that confirmed the cluster status or something else)
3. How did the authors end up with the final set of 6 (or 3?) scatter plots from which they have derived their property classes? Were “linear-clusters” identified only in these plots or did the authors select the final plots based on relevance of plotted variables in folded proteins. If latter was true – what was the criteria they used to identify the most relevant scatter-plots
4. All property classes including the classes #6, #7, and #8 should be precisely defined. The description “classes #6, #7 and #8 were derived from 49 fundamental amino acid properties and derived scales that are based upon an analysis with Analysis of Patterns (ANOPA)” is unacceptable. Which of 49 fundamental amino acid properties and their derived scales were used, and how, to identify property classes from #6 to #8.
5. Scatter-plots that were used for generation of classes from 5 to 8 should be shown.

5 - Comments on the reporting style
1. The Introduction section is quite short – the authors should elaborate more on relevant physico-chemical properties of amino acids and their importance in protein folding in this section.
2. There are parts of the Introduction in the Results section (the first paragraph) and the Discussion section (alignment matrices).
3. The hydrophobic scale that was chosen as the optimal one was normalized average of three published hydrophobicity scales that were found “most robust in correlation analyses” with robustness vaguely defined in the Methods section as associations to “multiple fundamental properties of the 20 natural amino acids using multi-variate statistical procedures, thermodynamics and biophysical chemistry considerations”. It is just latter, at the very end of the Results section that one can find out that “robust” scales are actually those whose MLR models using property classes as dependent variables exhibited highest R2 values. The Methods section should be written more clearly.
4. Table 2 – The labelling of Table 2 should be improved as authors keep explaining what is presented in which column of the Table 2 throughout the Results section.
**Competing Interests:** No competing interests were disclosed.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to state that I do not consider it to be of an acceptable scientific standard, for reasons outlined above.

Author Response 25 Aug 2020

**David Cavanaugh**, Benchmark Electronics, Huntsville, Alabama, USA

**Ana Jeroncic**

"... Ambiguously defined classes #6, #7 and #8 were derived from 49 fundamental amino-acid properties and derived scales that are based upon an analysis with Analysis of Patterns (ANOPA) ..."

- The 49 AA properties are represented by:
  - 2 sequence frequency scales
  - 6 secondary structure propensity scales
  - 8 hydrophobicity scales
  - 4 free energy scales (in water, protein folding/unfolding)
  - 9 HPLC retention time scales
  - 4 probability of an AA inside a folded protein core or on the outside
  - 7 molecular property scales
  - 9 physical (measured) property scales

- The ANOPA method is a pattern recognition, pattern projection method. The ANOPA procedure projects n-space pattern point/vectors into a 3 dimensional object which is a cylinder. The axis of the cylinder, which is termed the relation vector, is formed from the pattern point centroid (averages of each AA property for all points) to an outgroup average point (averages of each AA property for a pair of selected points). The outgroup pair are selected on the basis of a histogram of the pattern point Euclidian distances from the pattern point centroid. Two pattern point distances are calculated with respect to the relation vector from a projection of each point onto the relation vector yielding the distance along the relation vector and the distance from the relation vector. The angle of rotation of each pattern point projection onto the relation vector is calculated. Thus, a cylindrical coordinate system is formed which is then converted into rectilinear coordinates. The $X'$ and $Y'$ rectilinear points are formed from each pattern point's pattern projection distance times the cosine and sine respectively of the pattern projection angle of rotation. The $Z'$ point is simply the pattern projection intersection distance along the relation vector.

- There is a relatively flat and linear structure to the higher order correlation structure in the 49 AA property pattern space. The 3D ANOPA $X'$ coordinate has a strong linear relationship ($R^2 = 95.13\%$) with the angle of rotation $\theta$ pattern point/vector relation vector projection vectors. Similarly, the 3D ANOPA $Y'$ coordinate has a strong linear
correlation ($R^2=99.86\%$) with the d2 distances of the pattern points/vectors from the relation vector.

- Each of the 3D ANOPA coordinates have meaningful interpretations/associations with amino-acid properties. The $X'$ coordinates have 3 good linear series when scatter plotted with AA refractivity [179] and AA mass/(area/volume) [192], and it has 4 linear series with the Tanford hydrophobicity scale [197]. The AA refractivity scale has a strong linear relationship with the AA mass/(area/volume) derived scale, except that Glycine is off the regression line. The $Y'$ coordinates have 2 linear series when scatter plotted against the DNA/RNA numerically encoded (0-63) lexical table (UCAG, UCTG) averages and 3 linear series when scatter plotted against the AA property derived scale $2^*H-B-DB$. In the latter scale $H$ is our hydrophobicity proclivity scales, $B$ is a beta sheet proclivity scale derived from several published statistical scales and $DB$ is the double bend proclivity scale derived from several literature statistical scales. The $2^*H-B-DB$ scale reliably predicts (a number of proteins were sampled) the presence (or propensity) of alpha helices (value>0) and beta sheets (value<0) with sinusoidal trends (running average of 3) as seen with the primary sequence AA ordinal numbers plotted along the X axis. Since the $Y'$ coordinate has a strong relationship with the $2^*H-B-DB$ scale, which itself has a strong relationship with protein secondary structure we see that the $Y'$ coordinate also has a relationship with protein secondary structure. We also can infer that there is a definite relationship between protein secondary structure and the DNA code, which in itself is related to AA hydrophobicity that we have documented in the hydrophobicity paper and the fact that the secondary structure proclivity scale by definition is related to hydrophobicity. Finally, the 3D ANOPA coordinate $Z'$ strongly correlates with all of the AA hydrophobicity scales (and some of the AA HPLC scales) in our database, both those used in the 49 property ANOPA analysis and those that were not used in this analysis.

"... MLR $R^2$ value represents a rigorous test for a robust, high performance hydrophobicity scale ... The rationale for such an assumption was that for all tested hydrophobicity scales, the $R^2$ values of the Multiple Linear Regression (MLR) were higher than the $R^2$ values of the simple binary regression models ... I disagree with the authors on the MLR $R^2$ rationale as I have concerns about the appropriateness of the data analysis"

- The argument presented is not that the MLR regression was superior to that of the binary regression because of higher correlation coefficients, although possibly that may be true, but rather that the AA property class MLR represented more information in that it represented the behaviors of amino-acids in a larger series of contexts and properties. Also, the argument implied by having a MLR with the 8 property class scales is that for each context the amino-acids partition into sub-sets and that these different context sub-sets join together to determine amino-acid behavior in protein folding, interactions with water, interactions with membranes, secondary structure and electronic behaviors associated with AA-AA interactions. Moreover, we argue that within any given regression that the higher the correlation coefficient, the stronger the evidence for the superiority of a given hydrophobicity scale within that regression criterion. Top performance of a hydrophobicity scale within several regression relationship criterions based upon different property information leads to a performance/robustness conclusion based upon the consilience of the evidence.
We use the coefficient of determination ($R^2$) rather than the correlation coefficient because it is a much more conservative statistic than the correlation coefficient ($R$) and can meaningfully be interpreted as the percent linearity between the dependent variable and the $X$ independent variable(s) in the regression. Concerns might be raised about two points in the property class MLR, which are the possibility of inter-correlation between $X$ variables and the loss of degrees of freedom in the regression statistics (over-fitting). We deal with these concerns in several fashions. First off, we calculate a $T$ test significance only for the regression itself and not for any individual $X$ variable regression coefficient since the $Y$ variable regression performance itself is what is being measured. Secondly, the correlation coefficient $T$ test (null hypothesis being that $R=0$) number of degrees of freedom (20) is reduce by 2 for each $X$ variable coefficient in the regression and by 1 for the $Y$ axis intercept. In the AA property class MLR we calculate the resulting degrees of freedom in the regression and correlation coefficient $T$ test as $20-(2*8+1) =3$. We allow for some dimensionality reduction due to inter-$X$ pair correlation as long as the dimensionality reduction (i.e. percent reduction in the number of states with discrete variables) is not large and new information is being brought forth between each pair of $X$ variables. Where there is no new information adduced by inclusion of a $X$ variable we find that the magnitude of the correlation coefficients are reduced and significantly reduced by the squaring process to get the coefficient of determination. Furthermore, we demand that any set of property class vectors produce large significant coefficient of determinations with a large number of disparate AA property scales in our database. The latter criterion produces a very, very high degree of statistical confidence that the collection of property class scales can serve as a highly robust basis set for MLR vector regressions that can evaluate the robustness/significance of individual AA property scales.

"... the reporting in the manuscript should be substantially improved ..."

We agree and are revising the manuscript accordingly.

"Throughout the paper the description of the MLR models is very confusing ... it is not clear what models were actually run ... what was the dependent and independent variables ... what was the estimated regression coefficients and their statistical significance ..."

○ We have addressed these comments to some extent above. Additionally we will say that we will move all of the related discussion of regression methodology into the methods section where the discussion will make more sense. The key point is that the MLR $X$'s are not single variables, but rather are column vectors with each column vector having values for each amino-acid. In this context we equate the concept of a scale with the concept of a column vector. Having made this distinction, the concept of a Multi-Linear Regression (MLR) is still valid and simply represents a broader mathematical context. Our use of the MLR methodology with the 8 property class scales is to evaluate a correlation relationship as whole, so the entire regression correlation is what is used/important and the statistical significance of individual variable regression coefficients is meaningless. If the MLR method was being used to fit a variable for the purpose of extrapolating new $Y$ values with scale values outside of the basis set, such as with new amino-acids, then the statistical significance of
variable coefficients would be germane.
"eight property classes ... vector sets of clusters/linear families of curves in multiple linear regression relationships between two (or more) amino-acid physico-chemical properties .. which is very confusing ... seems like more independent variables are being added than the eight property class variables ..."

- One must keep in mind that the MLR method that we are using uses vectors (scales) of dimensionality of 20 (one for each AA) and not individual variables. When doing AA property scale pair scatter plots between each property scale selected for our database we often find clustering behavior which indicates that the amino-acids are partitioning into distinct sub-sets and each amino-acid can be assigned a numerical value for the ordinal number of the sub-set into which it partitions. The clustering relationships uncovered for any particular pair of two AA property scales is generally driven by molecular geometry, secondary group geometry, numbers and types of atoms/inter-atomic bonding in the secondary group, molecular size, molecular mass, molecular volume, molecular surface area and entropy distribution about the molecule. No given pair of AA property scale relationships represent the full range of these molecular properties and the nature of the interaction with water and cellular membranes. The key point though is that there are distinct sets and sub-sets, which when numerically encoded can jointly describe each amino-acid as a row vector of some dimensionality M. We have found through extensive analysis that M is a very good size. Within each property pair scatterplot where clustering occurs, we can have different patterns such as unstructured clusters, multiple quasi-parallel linear clusters or multiple linear clusters that intersect at some point (often at glycine or alanine). Generally, there is a geometric ordering that allows the assignment of a meaningful ordinal number. To reiterate, a property class scale (column vector) represents a relationship between property scales and the values assigned within the scale to each amino-acid are their respective sub-set/cluster ordinal number. Property class scales assembled in this way can be used as one of the basis set vectors to be used to evaluate the performance and reliability of any individual amino-acid property scale. We also need to note that the three property class vectors defined between the 3D ANOPA coordinate system planes represent the correlation based dimensionality reduction from a pattern space of dimension 49 to a pattern space of dimension 3.

"... majority of these property classes were actually ordinal variables ... more actual variables should have been added owing to the introduction of dummy variables ... sample size of these models too small to estimate model parameters precisely ... "

- In fact all of the numbers within the 8 property class scales are ordinal numbers representing real facts of clustering and the distinct geometrical ordering of the clusters. The practice of using assigned numbers to sets/sub-sets of objects, especially if the numbers can represent some form of ordering, is widely used in multi-variate statistical procedures such as MLR regression, discriminant analysis, Principal Component Analysis (PCA)and procedures cognate to PCA, as is found in fieldids such as numerical taxonomy, computational biology and other fields that numerically encode state data. The authors have enjoyed great success in using discrete data to represent state data in a number of contexts over a number of years.
Since the MLR regressions are not used for predictive purposes, but rather to express the degree of correlation and relatedness as expressed in the correlation coefficient and its T test based P value, the concept if regression coefficient statistical significance is a non-sequitur.

"... R^2 quite inflated ... over fitted model ... large number of independent variables that the authors did not adjust for compared to simple linear binary "regression models ..."

○ The authors would point out that additional variables do not inflate R^2 values unless there is a very serious inter-variable correlation between a large sub-set of the X variables. To the contrary, we find that working with the amino-acid property scales that we have assembled that there is generally an artificial deflation of R^2 relationships. Therefore, we do not agree that the introduction of variables into an MLR artificially inflates a MLR correlation versus a binary regression correlation. We adjust for these concerns in three ways: use of the coefficient of determination statistic is very conservative and we screen out a number of potentially significant relationships thereby; we use a T test to assess the statistical significance of the whole regression relationship and reduce the number of degrees of freedom of the T test accordingly as described above; we uncover the specific physical meanings reflected in each property scale relationship cluster to ensure that the physical-chemical factors at play are distinct and bring new information to the table.

"... problem of multicollinearity between independent variables ... inflated MLR R^2 ... table 2 PC2 and PC4 had a Kendall tau coefficient of 0.82, P<0.001"

○ We note that property classes 1 and 2 devolve from the relationship between the hydrophobicity proclivity scale and specific absolute entropy and that property classes 3 and 4 devolve from the relationship between the hydrophobicity proclivity scale and the specific volume as seen in figures 2 and 1 respectively. Both the specific absolute entropy and the specific volume are thermodynamic intrinsic property scales formed by the normalization of the amino-acid molecular volumes and molecular absolute entropy by their molecular masses, which create two secondary or derived scales from fundamental molecular properties that have the molecular mass in common. Furthermore, both the total entropy and the volume are related in some way to the molecular mass, so dividing by the molecular mass in both cases offset that dependence and leaves an intrinsic average property. Property class 2 partitions into two linear subgroups on the basis of polar and non-polar amino-acids. Property class 4 partitions into 4 linear subgroups on the basis of size, relative amounts of non-polar area and the presence/absence of a relatively strong or weak polar group. While property class 4 does reflect the polar/non-polar distinction to some extent, it does not reflect a strong and clear partition of AA by polarity or non-polarity, therefore new information is introduced by the addition of property class 4. We note that there are 2*4 possible combinations of states between property classes 2 and 4, but in fact there are actually 5 states formed giving a dimensionality reduction of 1-5/8 =37.5%, which we consider to be acceptable. We do not see strong enhancement of MLR correlation coefficients with any statistically significant AA property regression with the 8 property class scales.
interrelationship of fundamental molecular properties based upon AA molecular size, molecular geometry, entropy/entropy density, mass/mass density, surface area or electronic configuration/hybridization, some inter-property class correlation is inevitable and must be tolerated to make any progress in understanding amino-acid physico-chemical properties and protein folding. See the additional relevant discussion above.

"... main result of this paper ... based upon assumption that amino acid property classes forming linear clusters ... within context of folded proteins ... is not a valid assumption"

○ We agree that the 8 property class scales are one of the major results of this study, but we do not agree that it is the only one. We do not assume that the 8 amino-acid property class scales are a major determinate of protein folding per se, but rather we assume the fundamental molecular properties of each amino-acid in the context of a folded protein environment or the context of contact with a water environment are what determine protein folding. The assemblage of the 8 amino-acid property class scales that we report reflects a wide spread clustering pattern (sometimes in linear series) in 2D scatterplots and 3D ANOPA scatterplots between two or more scales. The 8 amino-acid property class scales we have assembled for our present study reflect the best of the clustering patterns forming unique subsets of amino-acids in different contexts/molecular properties with regard to hydrophobicity scales, but also with a number of other amino-acid physico-chemical property scales, secondary structure statistical propensity scales and molecular property scales. To the extent that any given hydrophobicity scales correlate with, and were determined from folded proteins, reflecting secondary/super-secondary structure of folded proteins, then the 8 property classes that we report are relevant to folded proteins. The 8 property classes that we report have statistically significant (P < 0.05) MLR regressions with 124 amino-acid property scales in our assembled database, which implies that the amino-acid clustering behaviors reflected in these numerical scales can largely be reproduced with the clustering behaviors of the amino-acids. While the idea of linking of hydrophobicity to protein folding is an important idea, the finding that nature has selected the 20 natural amino-acids on the basis of distinct sub-sets/clustering in several different joint property dimensions is a significant finding in and of itself, perhaps leading to criteria for the engineered selection of un-natural amino-acids that could provide unique properties to active enzymes/proteins.

"... R^2 is an overused statistic for linear regression analysis and additional metrics are required to get the whole picture ...R^2 actually represents the square of the Pearson correlation coefficient"

○ We do not agree with the contention that the R^2 statistic is trivial and meaningless from overuse. Rather we point out that R^2 is a more conservative statistic than is the correlation coefficient and rolls off much more quickly, as well as having an important interpretation as the percent linearity that directly indicates the relative scatter of the regression calculated Y’s with respect to the actual Y’s.

"... A reader should know precisely which scatter plots were screened for a ‘linear-cluster’
There were approximately 175 amino-acid property scales used for our study, of which maybe 15-20 were derivative scales prepared from ratios of fundamental molecular properties such as volume and mass, which would represent intrinsic versus extrinsic scales in the thermodynamical sense. About another 10-15 scales represent averages of literature reported amino-acid property scales. Please note that linear clusters were not the other patterns seen. There were some non-linear patterns and simple clusters for example. Also note that the scales selected represent a very wide range of amino-acid properties including secondary structure statistical propensity scales, fundamental molecular properties like surface area/mass/volume, bulk properties such as index of refraction/melting point/pK-C, HPLC retention times, free energy in water, solvent/water partitioning, average fraction occurring in proteins, average amino-acid burial in proteins/amino-acid exposure wet water in folded proteins, hydrophobicity scales, NMR parameters, Rf parameters, several literature Principal Component Analysis/Factor Analysis parametric scales and other miscellaneous property scales. Of the 175 amino-acid property scales (and about ~200 total scales) 124 scales had statistically significant (P < 0.05) MLR regression correlation with the 8 property classes that we have reported.

"... the eight amino-acid property classes most important novelty of this paper ... what is the reasoning that AA property classes represent real AA physico-chemical properties in the context of folded proteins ... rather was the assumption that high R^2 from MLR justifies assertion of relatedness to physico-chemical properties ... a high R^2 not a valid basis for assumption ..."

We agree that the 8 amino-acid property scales are novel and an important finding. We disagree with the comments on the MLR regression correlation and assert that high MLR R^2 values are both statistically and physically valid measures for reasons discussed above. The high and statistically significant coefficients of determinations are spread over many types of amino-acid physico-chemical property scales, including structural statistical propensity scales and hydrophobicity scales. We find that there is a strong relationship between amino-acid HPLC retention times and a number of the higher quality hydrophobicity scales in our assembled database, hence there is a direct link between a measureable amino-acid bulk property and amino-acid partitioning behavior in the structure of folded proteins.

"... the authors should describe the method that they used to identify linear clusters on a plot ... follow up analysis of physico-chemical/biochemical properties of clusters ... regression analysis ..."

We did describe most of the points commented here in the original manuscripts, but agree that the descriptions need to be expanded upon. There were a number of linear clustering patterns, amongst other patterns, that can be described as multiple quasi-parallel series, multi-linear series that intersect at a given amino-acid serving as a quasi or virtual origin (e.g. Karplus 1997), quasi-parallel cross hatched linear
patterns (e.g. figures 1 & 2 in the manuscript). We ran, although we did not report a regression correlation coefficient on each series to assess its relative linearity. Where we found what we discerned as significant relationships, we evaluated those putative relationships from the point of view of physical reality as expressed by fundamental molecular properties and potential relationships with water, especially with aqueous clathrate membranes.

"... how did the authors end up with the final set of scatter plots from which AA property classes were derived ... were they linear clusters ... did authors select scatterplots on basis of plotted variable relevance ... what was the criteria used to select most relevant scatterplots ..."

○ Our primary and starting premise was to look for single linear or non-linear patterns in all amino-acids in the scatterplot of any two given scales. Along the way we found that multiple linear series occurring in scatterplots between many pairs of amino-acid properties, which was a pattern we could not ignore. Many of these multi-family linear series ranged between clearly discernable to very high quality; the latter end of the range being what we concentrated on. We also cross checked to see if sister AA property scales in the same class such as hydrophobicity resulted in similar patterns, such as which we see in figures 1 and 2 and in table 2 of the manuscript. Once the linear/non-linear (single and multi-family) patterns were found a detailed review of these patterns were made to find underlying physico-chemical and biological reasons as well as statistical generalizations. The most relevant scatter plots were selected based upon their quality (visual and linear/non-linear regression) and explanatory power for protein structure and function and reliability/specificity for protein alignments.

"... all property classes including #6, #7 & #8 should be precisely defined ... derived from 49 fundamental amino-acid properties/derived scales ... Analysis of Patterns (ANOPA) ... need expanded description ..."

○ We have defined property classes #1 through #5, but we will expand upon these definitions. We will define property classes #6-#8 with plots and physical explanations. We will describe the ANOPA procedure and explain the specific 49 amino-acid property scale ANOPA analysis used in this study. The 49 amino-acid property scales used in the ANOPA analysis are property scales gleaned from the literature and no derivative scales were used for this portion of the overall analysis. We will be putting in a couple of new tables to define the 49 amino-acid properties used in the ANOPA analysis and the 124 statistically significant (by regression correlation) amino-acid property scales

"... show scatter plots for property classes #5 through #8 ..."

○ We agree to revise the manuscript accordingly.

"... need expanded introduction ... expand more on AA physico-chemical properties relevant
to protein folding ..."

- We agree to modify the manuscript according to these comments, except that we note that the scope of this paper is to define our hydrophobicity scale and its application to protein alignments, whereas we take up the challenge of applying this work to protein folding with an extensive analysis in our follow up paper “Hydrophobicity Revisited: a Molecular Story.” This latter manuscript is in an advanced state of preparation and can be provided for review, which is recommended since the treatment of the material in this manuscript is way beyond what we can put into the current manuscript you have reviewed. We will be putting in a couple of new tables to define the 49 amino-acid properties used in the ANOPA analysis and the 124 statistically significant (by regression correlation) amino-acid property scales.

"... there is material in the results section (first paragraph) and discussion section (alignment matrices) that belong in the introduction section ..."

- We will revise the manuscript per these comments.

"... hydrophobic scale chosen as optimal ... normalized average of 3 hydrophobicity scales ... most robust in correlation analysis ... robustness vaguely defined in methods section as association to multiple, fundamental AA properties using multivariate statistical procedures, thermodynamics and biophysical chemistry considerations ... later find out that this means $R^2$ values derived from MLR models ..."

- We have expanded upon what this means in our responses to your review. We will incorporate this material into the manuscript and significantly beef up the methods section to accommodate the spirit of your comments. We also note that the statistical correlations were only part of the rationale for defining a “robust” hydrophobicity scale, which is based upon bringing a coherent theoretical analysis to bear upon this work as well. We will be adding some additional material requested by Dr. Carter that should speak to these comments. We offer to make available for your review the TMATCH theory and application papers to provide additional justification for what constitutes a robust hydrophobicity scale as we use the term.

"The methods section needs to be written more clearly"

- We agree and will modify the manuscript accordingly.

"... Need better tag description for table 2 owing to the considerable discussion of table two columns in the manuscript ..."

- We agree and will modify the manuscript accordingly.

**Competing Interests:** No competing interests were disclosed.
Peer Review

This manuscript addresses a worthy problem: improving multiple sequence alignment via the use of enhanced amino acid similarity metrics would enhance our ability to draw inferences from sequences of proteins whose structures, were they known would establish homology, but which owing to divergence have unrecognizably homologous sequences. It seems almost certain that we should be able to do a better job at homology searches if more about how amino acid physical chemistry leads to protein structure. It was for this reason that I agreed to review this manuscript based on the abstract.

The authors allude to work they have done that demonstrates the value of the new scales they describe here, but there is essentially no coverage of this central question in this manuscript, which is disappointing and detracts substantially from the value of the paper.

The device advocated by the authors is a neologism they call a “hydrophobic proclivity index”. This index is the result of statistical modeling from a variety of different scales of what has been called “hydrophobicity” and their derivatives (with respect to which variables is not described) in order to maximize agreement between the scale and calculations of the exposure of each of the twenty amino acids in folded proteins. The resulting presentation is interesting and potentially relevant, but is deficient its citation of the literature, and in results indicating either their methods or the results to which they allude. I conclude that the although the work described is well-motivated, and may lead to better homology searches, it nevertheless suffers from a variety of methodological and conceptual problems that may in the end compromise the work quite seriously. These are summarized below.

The data base:

The quest for a single “predictor” for the degree to which each amino acid is exposed on average in folded proteins has a long history. The authors have cited just about every previous attempt to correlate the two variables, but have excluded the one set of experimental data representing the actual physical chemistry of the twenty amino acid side chains, the vapor to water and water to cyclohexane distributions of side chain mimics measured and re-measured by Wolfenden's group \[1-5\]. Wolfenden has argued persuasively that octanol is a very unsatisfactory reference solvent for a variety of reasons, in part because of the ability of side chains to bring variable amounts of bound water into it from aqueous solution.
Omitting the Wolfenden free energies is a grave oversight, because it means that the regression analyses they describe are looking for signal in a variety of data sets that have already been corrupted by similar unsuccessful attempts by previous investigators who have kludged the extant variety of multiple scales. For that reason, any useful result the present authors may have achieved is likely to be idiosyncratic and only indirectly based on physical chemistry. Moreover, the authors provide no evidence of statistical tests that might suggest significance, and the correlations they describe, some of which are more impressive than others, are very likely to be successful only in proportion to the number of parameters from which their models are built and, I suspect, of somewhat circular logic.

**Relating protein structure to amino acid physical chemistry is very probably multi-dimensional.**

It is very probable that the inability of previous researchers to arrive at a single scale that predicts the accessible surface area in folded proteins arises because the problem itself is multi-dimensional. The authors describe a variety of classification schemes derived from attempts to rationalize scatterplots of amino acid properties. Indeed, they mention that one useful additional classification is likely related to the size of the side chain.

**Recommendation:** The authors should read carefully the papers from Wolfenden and Carter in which those authors describe first the correlation between the free energies of vapor to water distribution coefficients and amino acid side chain volume, and second, their success in predicting Moelbert’s accessible surface areas using a two-dimensional coordinate system one axis of which is the free energies, respectively, of water to cyclohexane and vapor to cyclohexane partition coefficients.

In conclusion, what might be of interest in this paper is the TMATCH algorithm and the improvements it brings to homology searches. That is not described at all. Instead, there are a variety descriptions of how a multitude of idiosyncratic hydrophobicity scales describing amino acid physical chemistry, notably excluding the (only) authentic ones, might be combined into one that predicts exposed accessible surface area by an algorithm that essentially produces a linear combination that is correlated with ASA by hidden, but nevertheless circular reasoning.

**References**

1. Wolfenden R, Lewis CA, Yuan Y, Carter CW: Temperature dependence of amino acid hydrophobicities.*Proc Natl Acad Sci U S A.* 2015; 112 (24): 7484-8 PubMed Abstract | Publisher Full Text
2. Wolfenden R: Experimental measures of amino acid hydrophobicity and the topology of transmembrane and globular proteins.*J Gen Physiol.* 2007; 129 (5): 357-62 PubMed Abstract | Publisher Full Text
3. Gibbs P, Radzicka A, Wolfenden R: The anomalous hydrophilic character of proline. *Journal of the American Chemical Society.* 1991; 113 (12): 4714-4715 PubMed Abstract | Publisher Full Text
4. Radzicka A, Wolfenden R: Comparing the polarities of the amino acids: side-chain distribution coefficients between the vapor phase, cyclohexane, 1-octanol, and neutral aqueous solution. *Biochemistry.* 1988; 27 (5): 1664-1670 PubMed Abstract
5. Wolfenden RV, Cullis PM, Southgate CC: Water, protein folding, and the genetic code.*Science.* 1979; 206 (4418): 575-7 PubMed Abstract
6. Carter CW, Wolfenden R: tRNA acceptor-stem and anticodon bases embed separate features of amino acid chemistry.*RNA Biol.* 2016; 13 (2): 145-51 PubMed Abstract | Publisher Full Text
7. Carter CW, Wolfenden R: tRNA acceptor stem and anticodon bases form independent codes related to protein folding. Proc Natl Acad Sci U S A. 2015; 112 (24): 7489-94 PubMed Abstract | Publisher Full Text

**Competing Interests:** No competing interests were disclosed.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

---

**Author Response 25 Aug 2020**

**David Cavanaugh**, Benchmark Electronics, Huntsville, Alabama, USA

**Charlie Carter**

"... We should be able to do a better job at homology searches if more about how amino-acid physical chemistry leads to protein structure ... the authors allude to work they have done that demonstrates the value of the new scales they describe here, but there is essentially no coverage of this central question in this manuscript, which is disappointing and detracts substantially from the value of the paper"

- This original work is described in 3 manuscripts, including this present paper. The other two manuscripts (theory and applications/results) describing the TMATCH alignment algorithm are almost ready for publication. These three manuscripts were all part of a single monograph that we had decided to partition up owing to length and the fact that the physical chemistry considerations of the current manuscript would be inappropriate for a bioinformatics journal and we needed to get this manuscript published to support another published paper (also available for review). We will incorporate into the present manuscript an appropriate level description of the TMATCH algorithm and actual results from these two manuscripts.

- A fourth manuscripts which goes significantly more deeply into the biophysics and thermodynamics of hydrophobicity and the broader justification for our hydrophobicity proclivity scale from these two points of view. We would like to provide you a copy of these three other manuscripts, which are in an advanced state of preparation, in answer to a number of your points. We agree with you that more summary material needs to be brought into this manuscript.

- We point out the excellent linear relationship between our hydrophobicity proclivity scale and the dpv average residue depth (from water) scale in figure 5 of the manuscript. The dpv scale is derived from a suite of proteins with actual solved folded structures, hence dpv is in effect a protein structural description of hydrophobicity where the deeper on average that a residue is buried, the more hydrophobic it is. Contrast the dpv concept with the ideas of average solvent exposed are, average buried area, percent buried, etc. When the relationships in figures 3-5 (with supporting literature) are taken together, the hydrophobicity potential driving protein folding is suggested to be the traditional solvent-solvent partitioning model of protein folding where amino-acids partition between aqueous exposure and burial within the hydrophobic core of folded proteins. We believe that this traditional model
of the primary driving force for globular protein folding is apt, but needs some
updates and modifications that we develop in detail in the “Hydrophobicity revisited:
A Molecular Story“ manuscript. We suggest in the present manuscript that aqueous
clathrates form about the hydrophobic area of amino-acids in an unfolded state (a
structural feature) and hydrophobic surface area on folded proteins possess a
surface tension like a gas bubble in water. We are suggesting that the physics driving
the initial stages of protein folding is the same physics as the coalescing/condensing
of bubbles of non-polar species or gas in water are essentially the same physics
which cause the mutual attraction of residues based upon the residue hydrophobic
area with associated aqueous clathrate. We expand upon this hypothesis in
“Hydrophobicity Revisited: A Molecular Story.”

- Our aim was to try to develop a scale which reliably represented the central
hydrophobic tendency (average) as a central/first order effect that allowed a simple
but meaningful paired comparison of amino-acids for homology relationships.
Adding additional variables/properties would in our opinion have detracted from our
goals of simplicity of calculation and utilization of what we believe is the primary first
order effect driving protein folding.

"... this index is the result of statistical modeling from a variety of different scales of what
has been called 'hydrophobicity' and derivatives ... derivative variables not described ...
presentation is interesting and potentially relevant, but is deficient in its citation of the
literature and in results indicating their methods or results to which they allude ... well
motivated ... suffers from a variety of methodological and conceptual problems ...
"

- We agree that an expansion of what we meant by derivative variables should be
included in the manuscript
- The derivative variables concept referred to in the manuscript are primarily ratios of
molecular properties or directly measureable bulk properties of amino-acids. In this
way we derive normalized intrinsic thermodynamical properties versus extrinsic/bulk
properties. For example, we find that the ratio of amino-acid surface area to their
volume has a relationship to hydrophobicity.
- Other derivative variables we used could be exemplified by estimating the water
cavity volume/surface area using amino-acid volume and surface areas and some
assumptions about the aqueous cavity geometry.
- We will be expanding our reference list of amino-acid scales utilized to better define
the amino-acid property scales we utilized in addition to hydrophobicity scales.
- We will also include the work of Wolfenden et. al. into the manuscript.

"... the authors have cited just about every effort to correlate a single hydrophobicity
predictor with amino-acid surface exposure in folded proteins ... but have excluded an
important hydrophobicity metric derived from classical physical chemistry methods by
Wolfenden and co-workers using water/vapor and water/cyclohexane partition distributions
of amino-acid side chain mimics ... Wolfenden has persuasively argued that n-octanol is a
very unsatisfactory reference solvent for multiple reasons such as the ability of side chains
to drag non-repeatable amounts of water into the n-Octanol solvent ... "

"... this index is the result of statistical modeling from a variety of different scales of what
has been called 'hydrophobicity' and derivatives ... derivative variables not described ...
presentation is interesting and potentially relevant, but is deficient in its citation of the
literature and in results indicating their methods or results to which they allude ... well
motivated ... suffers from a variety of methodological and conceptual problems ...
"
Since this experimental approach is unique, we will add additional material to the manuscript to incorporate these considerations and the work of Wolfenden et. al. and the implications of the work.

○ We agree that the standardization of water concentration within a polar solvent (immiscible with water) capable of hydrogen bonding is difficult in addition to the uncontrolled/variable amount of water dragged into the organic solvent phase in solvent/water partition experiments. We cover these issues at length in “Hydrophobicity Revisited: A Molecular Story,” but we agree that these considerations should be covered to some extent in the present manuscript. We also point out that solvent/water concentration ratios used to calculate free energy of transfers from water to the non-polar phase as a hydrophobicity measure relevant to protein folding suffer a systemic error if the organic solvent is incapable of hydrogen bonding as is the hydrophobic “solvent like” core of folded globular proteins. Our hydrophobicity proclivity scale has no units, but rather is a normalized proclivity. We do show in figure 4 of the manuscript that our hydrophobicity proclivity scale is directly relatable to free energy of solvent-solvent transfer using a popular dGow (water to n-octanol) scale.

○ We also point out that characterizing the amino-acid secondary groups alone does not treat the important contribution of the free energetics of folding by peptide bonds, such as is accounted for by some researchers by treating the 20 amino-acids as guests in the center of tri-peptides in octanol to water free energy studies.

“... omitting the Wolfenden free energies is a grave oversight ... are looking for signal in a variety of data sets that have already been corrupted by similar unsuccessful attempts by previous investigators who have kludged the extant variety of multiple scales ... any useful result the present authors may have achieved is likely to be idiosyncratic and only indirectly based upon physical chemistry ...”

○ We will include and discuss the work of Wolfenden et. al. due to its novelty and the additional insight that it provides

○ We largely agree with these statements by Dr. Carter regarding the large profusion of hydrophobicity scales in the literature. Our criticisms are a bit more muted in that we recognize and discuss some of the difficulties in trying to reflect the wide range of amino-acid behaviors and the difficulties inherent in trying to define hydrophobicity as a experimentally measurable concept related to the driving forces of protein folding.

○ We have relied upon other experimentally measured amino-acid properties to cross check our hydrophobicity proclivity scale using methods such as Multiple Linear Regression (MLR) as can be seen in table 3. For example, we can draw a good linear correlation between certain hydrophobicity scales, such as our hydrophobicity proclivity scale, and amino-acid reverse phase (C_{18} column) HPLC retention times (other researchers have drawn this conclusion as well).

“... The authors provide no evidence of statistical tests that might suggest significance ...”

○ We will provide both an F test from the MLR software we used and a Student's T test of the MLR (8 amino-acid property class scales) correlation coefficient R. We will show
these results in a new table with amino-acid properties versus the R², T and F significances. The MLR is significant where both the T test and the F test are significant at an alpha of 0.05 or better (actual alpha = 0.05*0.05 = 0.0025 or better). These statistical significance tests are on the MLR results and not the individual MLR coefficients.

"... the correlations they describe ... are very likely to be successful only in proportion to the number of parameters from which their models are built ... suspect some circular logic ..."

- The derivative scales as discussed above were derived from fundamental molecular properties or experimentally measured, so there is no concern on that score.
- The row scale property relationships with the properties in columns 1, 2 and 4 of table two reflect paired comparisons are independent amino-acid scales.
- The eight property class scales in column three of table 3 reflected in MLR pair wise comparisons in column two are independent, although 4 of the amino-acid property class scales are also reflected in in columns 1, 2 and 4 of table 3. The other 4 amino-acid property class scales come from separate relationships as described in the manuscript. The last 3 amino-acid property class scales in table 4 derived from an ANOPA analysis are based upon 49 amino-acid property scales, none of which appear in the row property scales, thus, are independent.
- The 8 property class scales in table 4 result in statistically significant MLR relationships with about 50 amino-acid property scales not included in the ANOPA analysis of 49 amino-acid property scales, but the MLR back check of these 49 amino-acid property scales resulted in most of these 49 amino-acid property scales having statistically significant results.
- Overall, about 150 amino-acid property scales were evaluated in the work.

"... probable that the inability of previous researchers to arrive at a single scale that predicts the accessible surface area in folded proteins arises because the problem itself is multi-dimensional ... authors describe a variety of classification schemes derived from attempts to rationalize scatterplots of amino acid properties ... authors mention useful additional classification likely related to size of AA side chain ...

- We agree that there are multiple factors involved with the free energetics of protein folding and that the available amino-acid property scales and hydrophobicity scales measure different aspects of the relationship of amino-acids to each other and to water. We believe that we should be able to related fundamental molecular properties of amino-acids and water into a meaningful larger picture.
- In the “Hydrophobicity Revisited: A Molecular Story” manuscript we demonstrate that the controlling factors for our hydrophobicity proclivity energy scale (regression corrected dGow) are: #H-bonds, polar surface area, non-polar surface area, diameter of the side chain, length of the side chain, residue volume, polar area/volume and surface area/entropy. The MLR regression coefficient of determination is 99.995% and is F test significant at 6.4x10⁻¹⁵.
- However, the point of the present work is to find a single scale that measures a central tendency of “hydrophobicity” with the assumption that residue hydrophobicity
(the contrast with water) is the dominant relationship between amino-acids for the purposes of protein alignments.

"... Read Wolfendson and Carter ... AA free energies of vapor to water partition vs. side chain volume ... Wolfenden & Carter accurately predicts Moelbert ASA with 2D plot/linear regression with AA water/vapor and AA water/cyclohexane partition confidents ..."

- Agreed. We will cover this material at an appropriate level of detail given the objective of the present manuscript.

"... need description of TMATCH and homology searches ... use multitude of idiosyncratic hydrophobicity scales describing AA physical chemistry ... notably excluding the only reliable/authentic physical chemistry scale describing AA by Wolfendsen and Carter ... concern that the hydrophobicity proclivity scale produces a linear combination that correlates well with the Moelbert ASA scale by hidden, but still circular reasoning ..."

- We will provide the TMATCH theory and applications manuscripts to supplement the material available for review.
- We will also include an appropriate amount of material from these two TMATCH manuscripts into the current manuscript.
- The Juretic average, Rose percent buried and Neumaier X hydrophobicity scales are independent of each other and the Moelbert ASA scale, therefore there can be no circular or self-referential physical chemistry scale describing AA by Wolfendsen and Carter ...
- We believe we have demonstrated that our hydrophobicity scale embodies the central tendency (i.e. first order effect) of the hydrophobicity phenomenon.
- We believe that the manuscript improvements requested Dr. Jeroncic will obviate any additional concerns with circular reasoning/analysis.

"Seven references to papers where Wolfendsen and/or Carter are one of the authors"

- Agreed. Thanks.

**Competing Interests:** No competing interests were disclosed.
Author Response 04 Oct 2020

David Cavanaugh, Benchmark Electronics, Huntsville, Alabama, USA

I am one of the authors of this paper. I wanted to update some news. The revision 2 of this paper, which reflects the reviewer comments, is now in press for F1000research. There is another paper on the subject of hydrophobicity published in ChemRxiv that is the follow up paper to this one:

https://chemrxiv.org/authors/David_Cavanaugh/8853095

**Competing Interests:** No competing interests were disclosed.

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com