RESEARCH ARTICLE

A hydrophobic proclivity index for protein alignments [version 1; referees: 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 We have created a hydrophobicity scale and a matching dynamic programming algorithm called TMATCH (unpublished report) that is able to match proteins with remote homologs with similar secondary/tertiary structure, even with very low primary sequence matches.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.

Corresponding author: Krishnan Chittur (chittuk@uah.edu)

How to cite this article: Cavanaugh D and Chittur K. A hydrophobic proclivity index for protein alignments [version 1; referees: 1 approved with reservations, 1 not approved] F1000Research 2015, 4:1097 (doi: 10.12688/f1000research.6348.1)

Copyright: © 2015 Cavanaugh D and Chittur K. This is an open access article distributed under the terms of the Creative Commons Attribution Licence, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Grant information: The author(s) declared that no grants were involved in supporting this work.

Competing interests: We declare that there are no competing interests for DC or KKC that have influenced the content of this article.

**Introduction**
An understanding of the properties and functions of a protein or a nucleic acid often begins with a search of the sequence against databases of proteins (or nucleic acids) with known properties or functions. The fundamental assumption is that sequence leads to structure which in turn leads to an understanding of the function. Search algorithms have improved and continue to improve. Yet, with proteins in particular, it remains difficult to detect remote homologies in the so-called twilight zone where proteins have low percent sequence identities starting around 20–25% and descending to around 10–15%. We describe a hydrophobicity scale that is proving to be an excellent measure of sequence relatedness. A robust estimate of the hydrophobicity based sequence identity can be calculated directly from a global alignment score, which may be directly used in database searches. 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 our new alignment algorithm (unpublished report) through the enhanced information content of our hydrophobicity proclivity scale.

**Approach**
Hydrophobicity scales (metrics) as understood in the literature are generally divided into four categories, derived from

- Experimental physico-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.

Through extensive analysis (primarily using regression and scatter plots), 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 describing amino acid properties. We derived a number of additional metrics by differentiating metrics that were describing amino acid properties. We derived a number of additional metrics by differentiating metrics that were describing amino acid properties. We derived a number of additional metrics by differentiating metrics that were describing amino acid properties. We derived a number of additional metrics by differentiating metrics that were describing amino acid properties. We derived a number of additional metrics by differentiating metrics that were describing amino acid properties. We derived a number of additional metrics by differentiating metrics that were describing amino acid properties.

**Methods**
We arrived at our hydrophobic 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 hydrophobic scale, or combination of scales, that best represented the role of the different amino acids in proteins?

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 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) of 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 linear clusters of amino acids, each set of which is considered to be a “property class”.

Consider Figure 1 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 acid, 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 I and a value of 1 to the next line and so on. In the same Figure 1 we can see the formation of Property Class 2 which contains only two series. 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 2). The four property classes we identified respectively in the scatter plots shown as Figure 1 and Figure 2, 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’s (shown as ace3H) having an R squared >95%.

Property class #5 reflects a scatter plot between the delta G of burial of AA secondary groups’ (as Y) and the number of atoms in the respective secondary group’s, which resulted in 5 linear series. Each of the linear series numbers (0 through 4) for each AA forms the basis of property class #5. The multiple linear regression of the delta G of secondary group burial with number of secondary group atoms and property class #5 resulted in an R² of 98.1%. Property 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). Together PC #1 to #8 represents 8 X vectors (listed in Table 3) in the multiple linear regression reported in the third column of Table 2. The property class index vectors are shown in Table 3.

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’.
Figure 1. Hydrophobic Proclivities versus Area per specific volume of amino acids.

Figure 2. Hydrophobic Proclivities versus specific absolute entropy.
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 amino-acid 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.

We created a normalized average of the three key hydrophobicity scales (The index i=1 is from Tang, index i=2 is from Neumaier and the index i=3 is from the average of the collected scales in Juretic). 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)

\[
Hn(i,j) = \left[ \frac{H_j - \min(H_j)}{\max(H_j) - \min(H_j)} \right] 
\]

\[
Hb(i,j) = \frac{H_{i1} + H_{i2} + H_{i3}}{3.0} 
\]

The hydrophobicity scale as calculated using Equation 2 using the scales published in 2,9,10 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.

An example scale, derived from an analysis of 28 literature hydrophobicity metrics, possesses a strong linear relationship (R² = 0.959) with our normalized average of three hydrophobicity scales, that forms a hydrophobicity proclivity scale, has been published in 1.

**Results**

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 alkyl 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 by examining the normalized average of several popular hydrophobicity scales versus the probability of an amino-acid solvent-exposed area (SEA) greater than 30 (shown in Figure 3).

![Figure 3. Normalized average of several hydrophobic scales with Solvent Exposed area.](image-url)
Figure 3 shows that there is indeed a relationship between the hydrophobicity scale 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 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.

In Figure 4 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\cite{7,19}. 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 above the correlation is quite good at 85.9% linearity (coefficient of determination). The regression of these two scales is used to derive a fitted 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. Table of Regression Fitted Hydrophobic Proclivities.**

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

Figure 4. Hydrophobic Proclivities versus Structure F & P Gtow.
Table 2. Linear correlation between hydrophobicity scales and AA physico-chemical properties.

<table>
<thead>
<tr>
<th>H Scale</th>
<th>Moelbert ASA R²</th>
<th>F &amp; P C₂OH R²</th>
<th>8 AA Property Class R²</th>
<th>6 factor R²</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chothia</td>
<td>44.1%</td>
<td>58.6%</td>
<td>86.0%</td>
<td>88.6%</td>
</tr>
<tr>
<td>Kyte-Doolittle</td>
<td>61.9%</td>
<td>65.7%</td>
<td>97.6%</td>
<td>94.8%</td>
</tr>
<tr>
<td>Jannin</td>
<td>56.2%</td>
<td>68.3%</td>
<td>84.1%</td>
<td>80.7%</td>
</tr>
<tr>
<td>Juretic Avg</td>
<td>63.7%</td>
<td>69.2%</td>
<td>97.9%</td>
<td>94.9%</td>
</tr>
<tr>
<td>SEA &gt;30</td>
<td>70.7%</td>
<td>72.7%</td>
<td>92.5%</td>
<td>89.4%</td>
</tr>
<tr>
<td>Engelmann-Steltz</td>
<td>53.0%</td>
<td>72.8%</td>
<td>78.3%</td>
<td>87.4%</td>
</tr>
<tr>
<td>Eisenberg-Weiss</td>
<td>56.4%</td>
<td>76.1%</td>
<td>86.4%</td>
<td>71.3%</td>
</tr>
<tr>
<td>Rose Avg % buried</td>
<td>86.1%</td>
<td>81.7%</td>
<td>88.1%</td>
<td>86.8%</td>
</tr>
<tr>
<td>Hopp-Woods</td>
<td>71.7%</td>
<td>82.7%</td>
<td>69.2%</td>
<td>71.3%</td>
</tr>
<tr>
<td>Tang Q</td>
<td>86.0%</td>
<td>84.6%</td>
<td>96.7%</td>
<td>91.3%</td>
</tr>
<tr>
<td>avg 3H</td>
<td>84.7%</td>
<td>85.9%</td>
<td>99.3%</td>
<td>95.8%</td>
</tr>
<tr>
<td>Neumeierer X</td>
<td>90.2%</td>
<td>89.1%</td>
<td>97.6%</td>
<td>94.2%</td>
</tr>
<tr>
<td>F &amp; P del G C₂OH</td>
<td>85.3%</td>
<td>100.0%</td>
<td>94.1%</td>
<td>89.6%</td>
</tr>
</tbody>
</table>

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\textsuperscript{[20,21]}. In Figure 5 we show that there is a strong relationship (97% 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.

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).

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, 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 physico-chemical 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 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.

The amino-acid Accessible Surface Area (ASA) has long been suggested as a reasonably accurate proxy for hydrophobicity as is also seen in a related scale, the Solvent Exposed Area > 30 square angstroms. 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

---

**Figure 5. Hydrophobic Proclivities versus Structure based mean residue depth.**

---
#1 to #8) and fourth column (Property Class #1 to #4, AA area/ specific volume' and specific absolute entropy') 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, 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 $R^2$) in regression comparisons with amino-acid physico-chemical properties as our hydrophobicity proclivity scale.

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/encoded 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.

In Table 2 column three is the 8 sets of numbers (vectors), dubbed as property classes and are eight X vectors in the multiple linear regression relationships with the $R^2$ shown in the third column. These eight property class vectors can form multi-linear regression fits with very high $R^2$ with a large number of the physico-chemical properties 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 physico-chemical 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.

### Discussion

The great organizing principle embodied within the hydrophobicity proclivities (and implied by dp), 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 the most hydrophobic residue 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 probability of occurrence of the 20 primary amino-acids is not the same throughout the domain/kingdoms of life, so this mathematical formalism can cause issues for identifying and aligning homologous proteins.

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. “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.”

2. “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.”

3. “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, 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 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.

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.

Competing interests
We declare that there are no competing interests for DC or KKC that have influenced the content of this article.

Grant information
The author(s) declared that no grants were involved in supporting this work.

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.
References


18. Online. Solvent accessibility. [Online Data]. Bordo Table 2: Solvent Exposed Area > 30 square angstroms calculated from data taken from 55 proteins in the Brookhaven data base, coming from 9 molecular families: globins, immunoglobulins, cytochromes c, serum proteases, subtilins, calcium binding proteins, acid proteases, toxins and virus capsid proteins. Reference Source


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 hydrophobicity 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 R² 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 adjust 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.

I have read this submission. I 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.

Competing Interests: No competing interests were disclosed.

Charles Carter
Department of Biochemistry and Biophysics, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

Peer Review Oath: I will be an ambassador for open science. I have benefited substantively from open reviews on several previous occasions, so I believe in its value. I will endeavor to be constructive, while at the same time remaining true to my own scientific values.
**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 \(^1\text{-}^5\). 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 \(^1\text{-}^6,^7\) 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 Moellbert'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

I have read this submission. I 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.

Competing Interests: No competing interests were disclosed.