Theoretical Determination of Amino Acid Substitution Groups based on Qualitative Physicochemical Properties

Kristine Yu

Abstract
This paper introduces a novel method for theoretical determination of amino acid substitution groups. The method here involves making a binary matrix based on 48 qualitative physicochemical properties and calculating a substitution matrix based on this using dot products. Isolated groups with high scores are determined to be valid substitution groups and conserved groups are derived from these valid groups. 258 valid groups and 31 conserved groups are found.

Introduction

Amino acid substitution groups have long been studied because they are useful for comparing groupwise and consensus relationships between proteins. Methods for developing substitution groups have been both theoretically based (Jiminez-Montano & Zamora-Cortina 1981; Kidera et al. 1985; Taylor 1986; Smith RF and Smith TF 1990; Mocz 1995) and empirically based.

Empirically, methods for developing amino acid substitution groups have been based on substitution matrices (Dayhoff 1978), which comprehensively describe the frequencies of one amino acid replacing another and statistically based on a comprehensive collection of amino acid properties (Miyata et al. 1979). Wu and Brutlag have also developed a method using a conditional distribution matrix to find empirically conserved in large protein databases (Wu and Brutlag 1996).

Theoretically, methods have primarily been based on quantitative physicochemical properties rather than qualitative properties. Taylor, most notably, classifies amino acids on the basis of qualitative properties (Taylor 1986). In his classification, substitution groups are the collection of intersection and union of sets of amino acids organized by property in a Venn Diagram.

Recently, work on amino acid substitution groups has addressed the importance of context in developing meaningful amino acid substitution groups. Ioerger has investigated the selection of physicochemical properties based on contexts specific for a given amino acid as defined as a pattern of hydropathy in a window surrounding the given amino acid (Ioerger 1997). The question of the effect of contexts on amino acid substitution groups, and indeed, the effect of amino acid properties on protein structure in general, remains fundamental. How do we know which properties are most important for determining protein structure? Are some properties more important than others in different contexts? How do we weigh amino acid properties in our modeling of amino acid properties when they are complexly interrelated?

We present here a revisit to classifying amino acid substitution groups based on qualitative physicochemical properties that is versatile enough to serve as both a general and specific model for amino acid substitution. Based on the properties collected, we form a matrix describing the classification of all 20 amino acids for all these properties. Based on dot products between column vectors representing the classification of a given amino acid based on each of the included physicochemical properties, we determine the relatedness between amino acids, and thus calculate a matrix of scores of pairwise substitutions between amino acids. Substitution groups sufficiently isolated with scores over a certain threshold are considered valid, and thus we end up with a multitude of groups much like the multitude of groups found by Taylor. However, we do not impose hierarchical or Venn diagram restrictions on our groups, and thus our groups

allow for a more general set of contexts than Taylor's Venn diagrams. We also find a conserved set of substitution groups, a subset of valid groups.

Methods

Physicochemical Properties

We collected 48 qualitative physicochemical properties describing side chain structure and functional groups, optical properties, hydrophobicity/charge/acid-base-properties, and size (volume and side chain length). The complete list of these properties as well and how we characterized the amino acids based on these properties can be found in Figure BLAH. In order to be included as a physicochemical property, a property had to be something that could be characterized (or at least well-estimated) by theoretical analysis of the amino acid structures. For example, which reactions an amino acid participates in or a comparison of the entropy of formation of the amino acids are not properties we can predict well simply by looking at the amino acid structures, but we can characterize the hydrophobicity of an amino acid by looking at properties of its sidechain such as what functional groups it has.

If experimental data was available on characterizing the amino acids based on some property, we disregarded it in favor of characterizing the amino acids based on theoretical methods alone. Thus, although multitudes of hydrophobicity scales exist, we decided if and how hydrophobic an amino acid was based on its polarity, functional groups, and structure. Similarly, we classified the amino acids based on size and length using only knowledge about the structure of the side chains.

We did not allow gradation within our classification of amino acids based on a given property, i.e. either an amino acid had the property or it did not. However, we did allow for gradation by having graduated properties, e.g. the more selective "very hydrophilic" as well as the all-inclusive "hydrophilic."

Admittedly, due to the qualitative nature of our classification (and indeed, due to the very nature of classification), whether or not we decided an amino acid had a given property could be quite subjective. Certainly there can be no argument as to which amino acids are branched at the β-carbon, but properties describing hydrophobicity/acidity in particular are less clear-cut. Thus, we give an extended explanation of how we classified amino acids for the properties that involved more subjectivity at http://www.stanford.edu/~krisyu.

Property and Substitution Matrices

Based on our binary classifications of the amino acids on the 48 properties, we created the 48 × 20 property matrix in Appendix 1. Each row vector of the matrix represents the classification of all 20 amino acids for a given property and each column vector represents the classification of a given amino acid based on each of the 48 properties. If amino acid *j* had property *i*, then we assigned the input in position *i,j* of the property matrix to be 1. If amino acid *j* did not have property *i*, then we assigned the input in position *i,j* of the property matrix to be –1.

We then calculated the substitution matrix in Appendix 2 using the property matrix as follows:

$$Position_{sub}(i,j) = (column_{prop}(i) \bullet column_{prop}(j))/48$$

We divided by 48 in order to normalize the substitution scores; these scores are given in Appendix 3. Accordingly, an amino acid's score for substituting for itself was 1.00.

The higher the substitution score for amino acid *i* substituting for *j*, the more alike amino acid *i* is to *j*, and hence the more likely *i* is to substitute for *j*. This is because of the following: if *i* and *j* both have or both do not have a given property *p*, then the portion of the dot product from

p is 1, but if *i* has the property but *j* does not or vice versa, then the portion of the dot product from *p* is –1.

Determination of Valid and Conserved Substitution Groups

We had two criteria for a valid substitution group: a high substitution score for the substituting amino acid(s), i.e. a score over a certain threshold, and isolation of the substitution group, i.e. a separation between the substitution score of the amino acid(s) of the substitution group and surrounding scores. We defined a substitution with a score of 0.50 or higher to be valid since this score marks the threshold between amino acids being alike and dislike, and we required an effective separation of at least one property between a substitution group's score and the surrounding scores for the group to be valid. Thus, because of our requirement of isolation, when a cluster of amino acids substituted against a given amino acid with identical substitution scores, we considered the cluster as a unit. Thus, any 2, 3, or 4-tuple combination of RNDEQ, amino acids which all scored 0.58 against H, would not form a valid substitution group with H; only the entire cluster RNDEQ would. A complete list of valid substitution groups is in Appendix 4.

Within the set of valid substitution groups, we considered the substitution groups equally valid. Thus, our set of valid substitution groups becomes analogous to Taylor's equally possible union and intersection combinations (Taylor 1986). However, we also considered a subset of the valid substitution groups which we called conserved substitution groups.

Due to the symmetry of the matrix (arising from the commutativity of dot products), the score of amino acid *a* substituting for amino acid *b* is identical to the score of *b* substituting for *a*. However, a valid substitution of amino acid *a* against amino acid *b* does not imply a valid substitution of *b* against *a*. Since either *b* or *a* could belong to a cluster, *a* might belong to a clustered valid substitution group of *b* but *a* itself would not be a valid substitution for *b* because it would not be isolated. Similarly, *b* might belong to a clustered valid substitution group of *a* but not form a valid substitution group of *a* alone.

Accordingly, we defined a conserved substitution group to be a valid substitution group for which the substitution group is valid for more than one of its constituents. The 20 individual amino acids are thus excluded from this set. As an example of a conserved substitution group, suppose that *bcd* is a substitution group for *a*. Then the minimum condition for this substitution group *abcd* to be considered conserved is that either *abc* has to be a valid substitution group for *d*, *abd* has to be a valid substitution group for *c*, or *acd* has to be a valid substitution group for *b*. (Certainly *abcd* is conserved if it is valid for more than just one of its constituent amino acids.)

Results

From our property and substitution matrices based on qualitative physicochemical properties of amino acids, we found 258 valid substitution groups (Appendix 4) and 31 conserved substitution groups (Figure 1). The sizes of the valid substitution groups were distributed as follows: 60 of size 2, 51 of size 3, 40 of size 4, 37 of size 5, 24 of size 6, 19 of size 7, 16 of size 8, 7 of size 9, 3 of size 10, and 1 of size 11. The conserved substitution groups included 17 of size 2, 4 of size 3, 2 of size 4, 4 of size 5, 1 of size 6, 2 of size 7, and 1 of size 8.

2 AL RK NQ ND DE EQ KH ML MF MI FL FW FI PI ST IL WY
3 AIM/IMA DQH/QDH MLF/FML ILV/LIV/VIL
4 MIAV/IVMA ILVM/LIVM
5 RKHEQ/KRHEQ MFIAV/IVMAF ILVMA/LIVMA DENQH/EDQNH/QNEDH
6 MLFIAV/ILVMAF/LIVMAF
7 EDQNHKR/HKRNDEQ ILVMAFP/LIVMAFP/VILAMPF
8 PVLMAIGF/VILAMPFG/LIVMAFPG

Figure 1. List of conserved substitution groups.

Since the validation of our substitution groups is theoretically based, we can easily justify all valid substitution groups by referencing the property matrix (Appendix 1). For example, we can see that AS is a valid group primarily because both are small and short. They both have the properties small, very small, and short according to our property matrix. Less obviously, AS both do not have a number of properties such as aromaticity or a phenol group in the side chain and their both not having these properties also contributes significantly to the substitution scores for AS. The largest valid substitution group is MLFIAVCPGWN which consists of 11 amino acids. All of these amino acids are defined as hydrophobic in the property matrix except for G and N, and Y is the only hydrophobic amino acid according to the property matrix which is not included. Y is not part of the substitution group although all other hydrophobics are because Y is aromatic and M is not. G and N both never share any properties with M, but negatively share many properties, i.e. do not have many of the properties of the property matrix, especially since there are very few properties we include that M has, and so they become part of the substitution group because of their lack of properties in common with the other amino acids.

The conserved substitution groups include a number of groups not found in previous theoretical determination of substitution groups. NQ, DE, ST, RK,ILV, and ILMV have been found among the theoretical determinations of Jimenez-Montano and Zamora-Cortina, Taylor, and Smith & Smith (Jiminez-Montano & Zamora-Cortina 1981; Taylor 1986; Smith RF and Smith TF 1990). However, the rest of our conserved substitution groups have not been found in previous theoretical work. In empirical work, though, Wu and Brutlag (Wu and Brutlag 1996) also found DN, EQ and EHKQR to be conserved in the BLOCKS and HSSP databases. Thus, only 9 of the conserved groups we found have been described before.

Just as we could for the valid groups, we can justify our conserved groups by referencing the property matrix. MI, for example, is conserved because MI are hydrophobic and strongly hydrophobic, and M and I both do not have many properties that are present in the property matrix such as aromaticity. (ML and MF, while valid, are not conserved because F is aromatic and L is aliphatic). Our largest conserved group is PVLMAIGF which is a combination of hydrophobics and aliphatics.

Unlike most other work with substitution groups, our conserved groups include no singletons due to our definition of conservation. Indeed, what most sets our conserved groups apart from others is the presence of P and G in large groups when they are often regarded as singletons as well as the frequency of M appearing in the groups. P and G appear in large groups most likely because of our predominance of properties based on size (in which P figures) and hydrophobicity (in which G figures) and because we exclude G from hydrophobicity classification, thus building higher substitution scores for G because G is neither hydrophobic, nor is it hydrophilic, and thus G is alike to the amino acids that are not hydrophobic (i.e. hydrophilic) as well as the amino acids that are not hydrophilic (i.e. hydrophobic). In addition, because all of the properties carry the same weight, the singleton groupings of P and G, e.g. symmetric alpha carbon or alpha imino sidechain, lose their importance in the midst of so many other properties. M figures in many of our conservation groups most likely because of the lack

of distinctive physicochemical properties that M has (and thus the lack of such properties in our matrix) and because the properties included in our matrix that M has are properties that large groups of amino acids have.

Discussion

Because our work with amino acid substitution is entirely theoretical, our model of amino acid substitution may not necessarily follow the way amino acids are conserved empirically. In addition, because we work only with the properties of free amino acids, our model does not necessarily correspond to the properties of amino acids in the contexts of their particular environments in proteins. Finally, due to the general nature of our approach, effectively averaging the classification of the amino acids across many physicochemical properties, our resulting model of amino acid substitution may not be useful when applied to specific contexts.

However, because empirically conserved substitution groups have been found to be well-explainable on the basis of biochemistry (Wu and Brutlag 1996), we should not abandon approaching amino acid substitution from a theoretical viewpoint, and a general approach relies on a much larger set of data than a context-dependent approach and thus is more likely to generate meaningful data. Moreover, a context-dependent approach can easily be adopted from our general approach. Firstly, we can introduce bias into the property matrix itself by our selection of properties. If we are working in a context where hydrophobicity is supremely important, for example, then we can include more properties relating to hydrophobicity and more graduated properties introducing shades of hydrophobicity. We can also introduce bias by weighting the properties. Again, suppose that hydrophobicity is most important. Then we can multiply the row vectors pertaining to hydrophobicity by some factor to increase the weight of hydrophobicity in the calculation of the substitution matrix.

A set of conserved substitution groups can provide an alphabet to describe discrete protein motifs and provide the basis for a better model of consensus relationships in multiple sequence alignment (Wu and Brutlag 1996). In particular, the development of context-dependent substitution groups based on methods for general substitution groups, such as the matrix methods presented here, may provide a versatile model for consensus relationships between proteins in general as well as protein in specific contexts.

References

Dayhoff, M. O, Schwartz, R. M., and Orcutt, B. C. 1978. A model of evolutionary change in proteins. In *Atlas of Protein Sequence and Function*, Nat. Biomed. Research Foundation, pages 345-352.

Ioerger, T. R. The context-dependence of amino acid properties. *Proceedings of the Fifth International Conference on Intelligent Systems for Molecular Biology* 1997;157-166.

Jiminez-Montano, M. A., and Zamora-Cortina, L. 1981. Evolutionary model for the generation of amino acid sequences and its application to the study of mammal alpha-hemoglobin chains. In Proceedings of the Seventh International Biophysics Congress, Mexico City.

Miyata, T., Miyazawa, S., and Yasunaga, T. 1979. Two types of amino acid substitution in protein evolution. *J Mol Evol* 12:219-236.

Smith, R. F. and Smith, T. F. Automatic generation of primary sequence patterns from sets of related protein sequences. *Proc Natl Acad Sci USA* 87:118-122, 1990.

Taylor, W. R. The classification of amino acid conservation. *J Theor Biol* 119:205-218, 1986.

Wu, T. D. and Brutlag, D. L. Discovering Empirically Conserved Amino Acid Substitution Groups in Databases of Protein Families. *Proc Int Conf Intell Syst Mol Biol* 1996;4:230-40

Appendix

1. Property matrix.

2. Normalized matrix of substitution scores.

3. Distribution of normalized substitution scores.

4. List of valid substitution groups.

Property	Amino Acid																			AA with property
	A	R	N	D	C	E	Q	G	H	I	L	K	M	F	P	S	T	W	Y	V	
aromatic	-1	-1	-1	-1	-1	-1	-1	-1	1	-1	-1	-1	-1	1	-1	-1	-1	1	1	-1	HFWY
UV absorbance	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	1	-1	-1	-1	1	1	-1	FWY
single aromatic ring	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	1	-1	-1	-1	-1	1	-1	FY
heteroaromatic	-1	-1	-1	-1	-1	-1	-1	-1	1	-1	-1	-1	-1	-1	-1	-1	-1	1	-1	-1	HW
aliphatic	1	-1	-1	-1	-1	-1	-1	1	-1	1	1	-1	-1	-1	1	-1	-1	-1	-1	1	GAILVP
branched	-1	-1	-1	-1	-1	-1	-1	-1	-1	1	1	-1	-1	-1	-1	-1	1	-1	-1	1	ILTV
branched beta-carbon	-1	-1	-1	-1	-1	-1	-1	-1	-1	1	-1	-1	-1	-1	-1	-1	1	-1	-1	1	ITV
flexible	-1	-1	-1	-1	-1	-1	-1	1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	G
inflexible	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	1	-1	-1	-1	-1	-1	P
alpha imino	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	1	-1	-1	-1	-1	-1	P
hydroxyl	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	1	1	-1	1	-1	STY
hydroxyl straight chain	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	1	1	-1	-1	-1	ST
phenol	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	1	-1	Y
sulfur	-1	-1	-1	-1	1	-1	-1	-1	-1	-1	-1	-1	1	-1	-1	-1	-1	-1	-1	-1	CM
sulfhydryl	-1	-1	-1	-1	1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	C
amide	-1	-1	1	-1	-1	-1	1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	NQ
carboxyl	-1	-1	-1	1	-1	1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	DE
carbonyl	-1	-1	1	1	-1	1	1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	NDEQ
imidazole	-1	-1	-1	-1	-1	-1	-1	-1	1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	H
guanidino	-1	1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	R
amino	-1	1	-1	-1	-1	-1	-1	-1	-1	-1	-1	1	-1	-1	-1	-1	-1	-1	-1	-1	RK
symmetrical alpha-C	-1	-1	-1	-1	-1	-1	-1	1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	G
alkyl	1	-1	-1	-1	-1	-1	-1	-1	-1	1	1	-1	-1	-1	-1	-1	-1	-1	-1	1	AILV
achiral	-1	-1	-1	-1	-1	-1	-1	1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	G
2 chiral centers	-1	-1	-1	-1	-1	-1	-1	-1	-1	1	-1	-1	-1	-1	-1	-1	1	-1	-1	-1	IT
ionizable	-1	1	-1	1	1	1	-1	-1	1	-1	-1	1	-1	-1	-1	-1	-1	-1	1	-1	RDCEHKY
charged (pH 6.5-7)	-1	1	-1	1	-1	1	-1	-1	1	-1	-1	1	-1	-1	-1	-1	-1	-1	-1	-1	RDEHK
acidic	-1	-1	-1	1	-1	1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	DE
basic	-1	1	-1	-1	-1	-1	-1	-1	1	-1	-1	1	-1	-1	-1	-1	-1	-1	-1	-1	RKH
strong basic	-1	1	-1	-1	-1	-1	-1	-1	-1	-1	-1	1	-1	-1	-1	-1	-1	-1	-1	-1	RK
polar/hydrophilic	-1	1	1	1	-1	1	1	-1	1	-1	-1	1	-1	-1	-1	1	1	1	1	-1	RNDEQHKSTWY
very hydrophilic	-1	1	1	1	-1	1	1	-1	1	-1	-1	1	-1	-1	-1	-1	-1	-1	-1	-1	RNDEQHK
weak hydrophilic	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	1	1	1	1	-1	STWY
hydrophobic	1	-1	-1	-1	1	-1	-1	-1	-1	1	1	-1	1	1	1	-1	-1	1	1	1	ACILMFPWYV
very hydrophobic	1	-1	-1	-1	1	-1	-1	-1	-1	1	1	-1	1	1	-1	-1	-1	-1	-1	1	ACILMFV
weak hydrophobic	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	1	-1	-1	1	1	-1	PWY
H-bonding	-1	1	1	1	1	1	1	-1	1	-1	-1	1	-1	-1	-1	1	1	1	1	-1	RNDCEQHKSTWY
H-acceptor	-1	-1	1	1	1	1	1	-1	1	-1	-1	-1	-1	-1	-1	1	1	-1	1	-1	NDCEQHSTY
H-donor	-1	1	1	-1	1	-1	1	-1	1	-1	-1	1	-1	-1	-1	1	1	1	1	-1	RNCQHKSTWY
tiny	1	-1	-1	-1	-1	-1	-1	1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	GA
very small	1	-1	-1	-1	1	-1	-1	1	-1	-1	-1	-1	-1	-1	-1	1	-1	-1	-1	-1	GASC
medium small	-1	-1	1	1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	1	-1	1	-1	-1	1	VTNDP
small	1	-1	1	1	1	-1	-1	1	-1	-1	-1	-1	-1	-1	1	1	1	-1	-1	1	GASCVTNDP
large (bulky)	-1	1	-1	-1	-1	-1	-1	-1	-1	-1	-1	1	-1	1	-1	-1	-1	1	1	-1	KRFYW
long	-1	1	-1	-1	-1	1	1	-1	-1	-1	-1	1	-1	-1	-1	-1	-1	-1	-1	-1	KREQ
very long	-1	1	-1	-1	-1	-1	-1	-1	-1	-1	-1	1	-1	-1	-1	-1	-1	-1	-1	-1	KR
medium-long	-1	-1	-1	-1	-1	1	1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	EQ
short	1	-1	-1	-1	1	-1	-1	1	-1	-1	-1	-1	-1	-1	-1	1	1	-1	-1	-1	GASCT

Appendix 1. Matrix of amino acid properties. Shaded/non-shaded blocks of properties indicate properties that share a category, i.e. optical properties; positive scores of 1 are shaded as well.

	A	R	N	D	C	E	Q	G	H	I	L	K	M	F	P	S	T	W	Y	V	
A	1.00	0.13	0.38	0.29	0.63	0.21	0.29	0.75	0.21	0.71	0.79	0.17	0.71	0.58	0.63	0.50	0.29	0.33	0.17	0.75	A
R	0.13	1.00	0.42	0.42	0.25	0.50	0.50	0.13	0.58	0.17	0.25	0.96	0.33	0.29	0.17	0.29	0.17	0.38	0.29	0.13	R
N	0.38	0.42	1.00	0.75	0.50	0.58	0.83	0.38	0.58	0.33	0.42	0.46	0.50	0.38	0.50	0.63	0.58	0.46	0.38	0.46	N
D	0.29	0.42	0.75	1.00	0.42	0.83	0.58	0.29	0.58	0.25	0.33	0.46	0.42	0.29	0.42	0.46	0.42	0.29	0.29	0.38	D
C	0.63	0.25	0.50	0.42	1.00	0.33	0.42	0.46	0.42	0.42	0.50	0.29	0.67	0.46	0.42	0.63	0.42	0.38	0.38	0.46	C
E	0.21	0.50	0.58	0.83	0.33	1.00	0.75	0.21	0.58	0.25	0.33	0.54	0.42	0.29	0.25	0.38	0.25	0.29	0.29	0.21	E
Q	0.29	0.50	0.83	0.58	0.42	0.75	1.00	0.29	0.58	0.33	0.42	0.54	0.50	0.38	0.33	0.54	0.42	0.46	0.38	0.29	Q
G	0.75	0.13	0.38	0.29	0.46	0.21	0.29	1.00	0.21	0.46	0.54	0.17	0.54	0.42	0.54	0.50	0.29	0.25	0.08	0.50	G
H	0.21	0.58	0.58	0.58	0.42	0.58	0.58	0.21	1.00	0.25	0.33	0.63	0.42	0.38	0.25	0.46	0.33	0.54	0.46	0.21	H
I	0.71	0.17	0.33	0.25	0.42	0.25	0.33	0.46	0.25	1.00	0.92	0.21	0.75	0.63	0.58	0.29	0.42	0.38	0.21	0.88	I
L	0.79	0.25	0.42	0.33	0.50	0.33	0.42	0.54	0.33	0.92	1.00	0.29	0.83	0.71	0.67	0.38	0.33	0.46	0.29	0.88	L
K	0.17	0.96	0.46	0.46	0.29	0.54	0.54	0.17	0.63	0.21	0.29	1.00	0.38	0.33	0.21	0.33	0.21	0.42	0.33	0.17	K
M	0.71	0.33	0.50	0.42	0.67	0.42	0.50	0.54	0.42	0.75	0.83	0.38	1.00	0.79	0.67	0.46	0.33	0.54	0.38	0.71	M
F	0.58	0.29	0.38	0.29	0.46	0.29	0.38	0.42	0.38	0.63	0.71	0.33	0.79	1.00	0.54	0.33	0.21	0.67	0.58	0.58	F
P	0.63	0.17	0.50	0.42	0.42	0.25	0.33	0.54	0.25	0.58	0.67	0.21	0.67	0.54	1.00	0.38	0.33	0.46	0.29	0.71	P
S	0.50	0.29	0.63	0.46	0.63	0.38	0.54	0.50	0.46	0.29	0.38	0.33	0.46	0.33	0.38	1.00	0.79	0.50	0.50	0.33	S
T	0.29	0.17	0.58	0.42	0.42	0.25	0.42	0.29	0.33	0.42	0.33	0.21	0.33	0.21	0.33	0.79	1.00	0.38	0.38	0.46	T
W	0.33	0.38	0.46	0.29	0.38	0.29	0.46	0.25	0.54	0.38	0.46	0.42	0.54	0.67	0.46	0.50	0.38	1.00	0.75	0.33	W
Y	0.17	0.29	0.38	0.29	0.38	0.29	0.38	0.08	0.46	0.21	0.29	0.33	0.38	0.58	0.29	0.50	0.38	0.75	1.00	0.17	Y
V	0.75	0.13	0.46	0.38	0.46	0.21	0.29	0.50	0.21	0.88	0.88	0.17	0.71	0.58	0.71	0.33	0.46	0.33	0.17	1.00	V
	A	R	N	D	C	E	Q	G	H	I	L	K	M	F	P	S	T	W	Y	V	

Appendix 2. Normalized matrix of substitution scores. The valid substitutions, defined to have a score 0.50 and above, are shaded in grey.

	----------Normalized substitution scores of valid substitutions----------																									No. of valid substitutions
AA	1.00	0.98	0.96	0.94	0.92	0.90	0.88	0.85	0.83	0.81	0.79	0.77	0.75	0.73	0.71	0.69	0.67	0.65	0.63	0.60	0.58	0.56	0.54	0.52	0.50	
A	A										L		GV		IM				CP		F				S	10
R	R		K																		H				EQ	5
N	N								Q				D						S		EHT				CMP	10
D	D								E				N								QH					5
C	C																M		AS						NL	6
E	E								D				Q								NH		K		R	7
Q	Q								N				E								DH		KS		RM	9
G	G												A										LMP		SV	7
H	H																		K		RNDEQ		W			8
I	I				L		V						M		A				F		P					7
L	L				I		V		M		A				F		P						G		C	9
K	K		R																H				EQ			5
M	M								L		F		I		AV		CP						GW		N	11
F	F						ACPF				M				L		W		I		AYV		P			9
P	P						ACPFS								V		LM		A		I		GF		N	9
S	S										T								NC				Q		AGY	8
T	T										S										N					3
W	W												Y				F						HM			5
Y	Y												W												S	4
V	V						IL						A		MP						F				G	8

Appendix 3. Distribution of normalized substitution scores of valid substitutions. Notice that for all amino acids, the scores of valid substitutions are either identical or separated by at least one normalized score. For example, for N: Q and D are 4 normalized scores apart; D and S are 6 normalized scores apart; EHT with all identical scores form a cluster 1 normalized score apart from S; CMP with all identical scores form a cluster 4 normalized scores apart from EHT.

----------Number of amino acids belonging to substitution group----------

1	2	3	4	5	6	7	8	9	10	11
A	AL	PI	AGV	FLW	ALGV	PGFN	AGVIM	FIAYV	ALGVIM	AGVIMCP	ALGVIMCP	AGVIMCPFS	ALGVIMCPFS	MLFIAVCPGWN
R	AF	PN	AIM	FWI	ACPF	STNC	AIMCP	FAYVP	AIMCPF	AIMCPFS	AGVIMCPF	NDSEHTCMP	NQDSEHTCMP	
N	AS	ST	ACP	PLM	RHEQ	SNCQ	ACPFS	PVLMA	NDSEHT	NEHTCMP	NSEHTCMP	QNEDHKSRM	MFIAVCPGWN	
D	RK	SQ	AFS	PAI	NQDS	SAGY	RKHEQ	PLMAI	CMASNL	EDQNHKR	QEDHKSRM	FMLWIAYVP		
C	RH	TS	RKH	PGF	NEHT	GASV	NSEHT	PAIGF	EDQNHK	QNEDHKS	MLFIAVCP	MIAVCPGWN		
E	NQ	TN	REQ	SNC	NCMP	GLMP	DENQH	STNCQ	EQNHKR	QDHKSRM	MIAVCPGW	PVLMAIGFN		
Q	ND	GA	NQD	TSN	DNQH	WFHM	CASNL	SQAGY	QEDHKS	MFIAVCP	MAVCPGWN	LIVMAFPGC		
G	NS	HK	NDS	GSV	CMAS	ILVM	EDQNH	GALMP	MLFIAV	MAVCPGW	FMLWIAYV			
H	DE	HW	DEN	ILV	EQNH	IVMA	EQNHK	ILVMA	MIAVCP	FLWIAYV	FLWIAYVP			
I	DN	IL	DQH	IVM	ENHK	IMAF	ENHKR	IVMAF	MCPGWN	FWIAYVP	PVLMAIGF			
L	CM	IV	CAS	IMA	QEDH	IAFP	QNEDH	IMAFP	FWIAYV	PLMAIGF	PLMAIGFN			
K	ED	IM	CNL	IAF	KHEQ	LIVM	QDHKS	LIVMA	FIAYVP	SNCQAGY	STNCQAGY			
M	EQ	IA	EDQ	IFP	MLFI	LVMA	QKSRM	LVMAF	PVLMAI	HKRNDEQ	HKRNDEQW			
F	EK	IF	ENH	LIV	MIAV	LAMF	KRHEQ	LMAFP	PAIGFN	HRNDEQW	VILAMPFG			
P	ER	IP	EKR	LVM	MGWN	LAFP	MFIAV	LAFPG	GLMPSV	ILVMAFP	LIVMAFPG			
S	QN	LI	QNE	LMA	FMLW	LFPG	MAVCP	LFPGC	HRNDEQ	LIVMAFP	LVMAFPGC			
T	QE	LV	QDH	LAF	FLWI	LPGC	MCPGW	WYFHM	ILVMAF	LVMAFPG				
W	KR	LM	QKS	LFP	FAYV	VILA	FMLWI	VAMPF	IVMAFP	LMAFPGC				
Y	KH	LA	QRM	LPG	PVLM	VAMP		VMPFG	LIVMAF	VILAMPF				
V	ML	LF	MLF	LGC	PLMA	VMPF			LVMAFP					
	MF	LP	MFI	WYF					LMAFPG					
	MI	LG	MAV	WHM					LAFPGC					
	MN	LC	MCP	YWS					VILAMP					
	FM	WY	MGW	VIL					VAMPFG					
	FL	WF	FML	VMP										
	FW	YW		VFG										
	FI	YS												
	FP	VA												
	PV	VF												
	PA	VG												

Appendix 4. List of valid substitution groups, organized by number of amino acids per group and alphabetically.