Elizabeth Villa-Rodriguez
PHYS498BIN

# Scoring matrices: Yet another approach.
# A review on Amino Acid similarity matrices based on forced fields by Z. Dosztanyi and A.E. Torda (1)

Scoring matrices are a powerful tool for today's biology. They are used on the macroscopic level, to construct phylogenetic trees and build evolution models, and at the molecular level, to predict protein's structure and function. Scoring matrices are used as a measure of how similar are two different amino acids. Many different methods have been developed to calculate scoring matrices, since the result of any analysis using them may be harmfully biased by the matrix selected.

The most wide spread matrices are the ones developed initially by Dayhoff in the 1970s (2). She made an extensive study of the frequencies in which amino acids substituted for each other during evolution. The studies involved carefully aligning all of the proteins in several families of proteins and then constructing phylogenetic trees for each family. Each phylogenetic tree was examined for the substitutions found on each branch. This lead to a table of the relative frequencies with which amino acids replace each other over a short evolutionary period.

There have been several attempts to construct more modern scoring matrices based on observed amino acid substitutions (3,4). The authors of these scoring matrices all report improved results compared to the original PAM matrices created by Dayhoff. There is no doubt that there is room to improve on the original Dayhoff matrices, both by including more data and by improving the underlying evolutionary model and mathematical methods. However ,it is not clear how much of an improvement they actually represent.

Totally different approaches have also been introduced. One avoids sequence similarity and considers structural similarity(5,6). Sequences can be aligned if the know three-dimensional structure is similar, even when their sequence have very low sequence similarity.

Other matrices, widely used these days, are based one pure physico-chemical properties of the amino acids, such as hydrophobicity, volume, composition and secondary structure preferences. It has been demonstrated that the first two properties are the most important determinants in the folding of a protein, and thus, the dominant factors underlying substitution in evolution (7,8).

In their recent paper, Dosztanyi and Torda (1) claim that using the information in molecular mechanics fore field leads to a more reliable mutation matrix, since these fields contain a natural weighting of different physical contributions. "A force field's parameters exactly define a weighting of everything from the Lennard-Jones parameters to bond angle constants".

This approach seemed very promising at the beginning of the paper, because they had other papers considering field forces for protein folding that had demonstrated to be successful (or, as successful as it gets).

The authors label a site of a protein by its native residue. After that, they perform a computational mutation of the wild amino acids by every one of the other 19 amino acids and record a score (energy) function that provides a direct measurement of the compatibility of every type of amino acid at that position. Furthermore, they average over all corresponding energy values where that particular amino acid is the native residue to obtain an element of the substitution matrix.

The calculations are based on low-resolution force fields (approximation and amino acid by a point mass), using knowledge-based score functions built for protein recognition (9). Their claim is that the results are force field independent, and for this they also calculate an amino acid substitution matrix built from a Boltzmann-based potential of mean force, taken from the literature. Since they are using additive force fields, the total score is decomposed into contributions from individual residues:

$$m_{AB} = 1/K_a \sum_{a=1}^{K_a} E_{a,B} \quad \ldots(1)$$

which reprensents a residue of type B in a site originally occupied by a residue of type A. There are $K_a$ sites with native residue A. This results in an asymmetric matrix. Assuming a normal distribution, they give an estimate of the reliability with the standard deviation:

$$\sigma_{A,B} = \sqrt{\frac{\sum_{i=1}^{K_a}\left(E_{a,A} - \left\langle E_{a,A}\right\rangle\right)^2}{K_a}} \qquad \ldots(2)$$

where summation runs over all $K_a$ a sites of native type A and $\left\langle E_{a,A}\right\rangle$ is equal to $m_{AB}$ by definition.

The sole data collection process provides some information about the force fields (maybe, more useful than the matrices themselves). The curves shown in figure 1 are show that the distribution of the energies are well approximated by a Gaussian, except for the Cys-Cys that introduces two peaks without previous knowledge the existence of two kinds of Cysteins ( disulfide and non-disulfide bonded).The standar deviation has a strong correlation with the absolute size of an energy. The distribution of the energies tend to be larger for hydrophobic residues, especially aliphatic examples. They provide a physical explanation: "the aliphatic hydrophobic residues may habe the most neighborurs and the environment with the greater variability".

The authors compare their obtained matrices with the most popular ones in the literature (8) and they found large similarity between them. It is both surprising and worth noting that the force field based matrices are so similar to the literature examples, despite that the later ones are made counting interchanges of residues in a set of sequentially or structurally aligned pairs, and the former ones use no alignments whatsoever in their construction process. One of the matrices is closer to the evolution-related matrices and the other is closer to the structural-based matrices.

The matrix based on the force field implemented by the sausage program (the "knowledge-based" force field) is similar to hydrophobicity scales, it accounts much better for hydrophobicity changes than most of the other matrices, but it disregards any other property such as volume, which is considered very important (!) The other matrix, worked very well for sequence alignments.

In my opinion, it is worth noting the introduction of these force field based matrices, but I do not think they provide any new information or approach in general terms. The authors themselves answer to the question how well these new matrices work with the known "the best matrix is problem specific".

So, are these matrices good at all or not? The great advantage of the new method relies in not using alignments for the construction of the matrix. This is very important since for making an alignment, one needs a scoring matrix in the first place!

Another interesting characteristic of this study, is the energy distributions obtained for each substitution, and the fact that they approach Gaussian curves with a characteristic width. I am not aware of anybody introducing this analysis before, and I am sure it will be useful for further analysis.

A third good thing about this particular kind of scoring matrix is that since they have less obvious connection with evolution, one can also use them for the studies of function, and not only structure of proteins, since one would expect the force fields not to be dependent on specific physico-chemical properties. The widely used scales mostly favor the measurement of hidrophobicity and volume, and these are not necessarily the most relevant features in the function of a protein. Sadly, this is not true. The sausage force field markedly shows a strong weighting of hydrophobicity over all other properties, including even volume!

Another inconsistency I am totally uncomfortable with, is the claim that the matrices are independent of the force field chosen. If the reader has been attentive, he/she must have noticed by now that I have been describing "the matrices" instead of "the matrix", and noted that one matrix is similar to the ones derived by physico-chemical properties while the other is quite useful for comparing distantly related sequences. This is not surprising since the two force fields are based on different construction philosophies which emphasize different aspects of sequence to structure alignment and fold recognition. In this sense, this approach is no good since again there will be a decision of the developer of the matrix of what is important to be considered.

The scoring matrices used today are very rooted in the bioinformatic tools. Unless some new revolutionary approach emerges which brings something very new to the calculations, the later will prevail.

I think it is possible to use this direction to develop better matrices, looking for a better way to characterize the force field, it might lead to a more general matrix that might be used for many problems, or at least less biased matrices for different problems.

## References

1. Dosztanyi, Z. and Torda AE.(2001) Amino acid similarity matrices based on force fields. Bioinformatics, Vol 17, 686-699.
2. Dayhoff M.O., Schwartz, R.M, and Orcutt, B.C. (1978) A model of evolutionary change in proteins. Matrices forf detecting distant relationships. In Dayhoff, M.O (ed), Atlas of Protein Sequence and Structure, Vol. 5, Sup. 3, National Biomedical Research Foundation, Washington SC, pp.345-358.
3. Gonnet,G.H., Cohen,M.A. and Benner,S.A. (1992) Exhaustive matching of the entire protein sequence database. Science, 256, 1443–1445.
4. Jones,D.T., Taylor,W.R. and Thornton,J.M. (1992b) The rapid generation of mutation data matrices from protein sequences. Comput. Appl. Biosci., 8, 275–282.
5. Johnson,M.S. and Overington,J.P. (1993) A structural basis for sequence comparisons. An evaluation of scoring methodologies. J. Mol. Biol., 233, 716–738.
6. Risler,J.L., Delorme,M.O., Delacroix,H. and Henaut,A. (1988) Amino acid substitutions in structurally related proteins. A pattern recognition approach. Determination of a new and efficient scoring matrix. J. Mol. Biol., 204, 1019–1029.
7. Grantham,R. (1974) Amino acid difference formula to help explain protein evolution. Science, 185, 862–864.

8. Kawashima,S. and Kanehisa,M. (2000) AAindex: amino acid index database. Nucleic Acids Res., 28, 374.
9. Huber,T. and Torda,A.E. (1998) Protein fold recognition without Boltzmann statistics or explicit physical basis. Protein Sci., 7, 142–149.

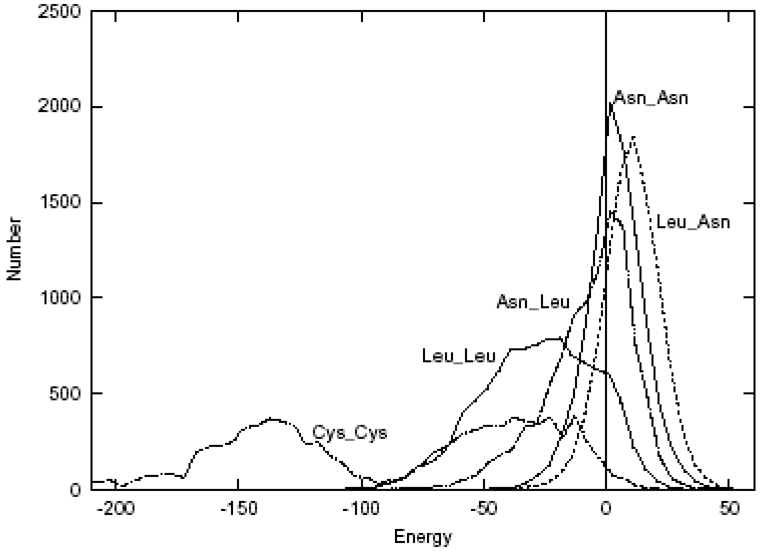

**Fig. 1.** Example for the energy distributions $E_{a,A}$ using the sausage force field. The examples are given for three native energy distributions (Cys_Cys, Leu_Leu and Asn_Asn) and for the replacement of Leu with Asn (Leu_Asn) and Asn with Leu (Asn_Leu). The more negative energy represents more favourable interactions.