Evaluation of protein hydropathy scales

Satu Jääskeläinen[1,2], Pentti Riikonen[1], Tapio Salakoski[1,2] and Mauno Vihinen[3,4]
[1]Department of Information Technology, University of Turku, FI-20014 Turku, Finland
[2]Bioinformatics Laboratory, Turku Centre for Computer Science, FI-20520 Turku,Finland
[3]Institute of Medical Technology, FI-33014 University of Tampere, Finland
[4]Research Unit, Tampere University Hospital, FI-33520 Tampere, Finland
sahaja@utu.fi, penrii@utu.fi, tapio.salakoski@utu.fi, mauno.vihinen@uta.fi

Abstract

Hydropathy is a dominant force in protein folding and has been measured with numerous methods. Several hydropathy scales are widely used in sequence-based predictions, however, without knowledge about their reliability. We investigated the prediction accuracy of 56 hydropathy scales by correlating predicted values with the accessible surface area in known 3-D structures of proteins. The correlations for the best scales are in the order of -0.26, whereas the weakest have on average merely -0.11. Results for different amino acids vary greatly within the scales, but are more consistent between the scales. One of the most common applications of hydropathy scales is to predict antigenic regions. Our analysis indicated that some epitopes are located among the most exposed regions. Despite poor overall correlation, hydropathy predictions can still be used in certain applications where qualitative analysis is sufficient.

1. Introduction

Hydropathy is a general term which refers both to hydrophobicity and hydrophilicity, the tendency of proteins and amino acids to like or dislike water interaction. Hydropathy has a profound effect on numerous protein features and is the dominant force in protein folding [1]. Several lattice based methods have been developed to predict the folding process based on hydropathy. Hydrophobic interactions in the protein core remarkably stabilize the structure. Hydrophobic interactions are energetic factors favoring the partitioning of an amino acid side chain from aqueous solution into a protein core or into a membrane bilayer. Hydropathy of proteins and amino acids has been investigated for a long time and several hydropathy scales have been produced for amino acids. Hydrophobic residues are more likely buried in the protein core than hydrophilic amino acids, which are more common on protein surface.

The more buried a residue, the more likely that mutations of the site will be pathogenic [2]. Cavity forming mutations in the hydrophobic core of T4 lysozyme decrease the stability [3] and mutations introducing charge into the core decrease the stability and enzyme activity. Hydropathy predictions have been applied to a number of protein properties. The predictions are easy to make - only the amino acid sequence and propensity table for the residues is needed. Hydropathy scales have been used extensively e.g. to membrane topology prediction, signal peptide recognition, prediction of regions on surface or core of the protein, epitopic regions, amphipathy of helical structures, protein secondary structures, turns, membrane-associated sequences, and surface β-strands. The property has also been applied to structural homology search by hydropathy profile alignment [4], and identification of novel membrane proteins [5]. Long hydrophilic runs are common only in membrane proteins [6].

Complementarity in hydropathies is important for several protein-protein and peptide interactions [7], and it can be used for protein docking [8].

Scales to measure and indicate amino acid hydropathy have been introduced since the 1970's. Currently there are a large number of scales available and, because the scales have been produced based on different measurements and principles, they often disagree on the degree of hydrophilicity/-phobicity of individual residues. This makes it difficult for prediction method users to select a suitable hydropathy scale. Since the calculation is always done with the same algorithm, the differences originate from the hydropathy scales. Here, we have performed systematic analysis of the accuracy of the hydropathy scales to predict the distribution of amino acids between the protein surface and interior, by correlating predictions with solvent accessible surface

as determined from three dimensional structures of proteins.

2. Results and Discussion

Protein hydropathy predictions are widely used for numerous purposes. However, there has not been information available about the accuracy of the methods apart from prediction of membrane topology (e.g. [9,10]) and surface β-strands [11]. Here we performed a systematic analysis of the prediction accuracy by comparing hydropathy predictions with experimental information, namely the accessible surface of the residues in proteins for which the three dimensional structure has been determined either by X-ray crystallography or NMR spectroscopy.

The adjustable parameters in hydropathy predictions are the prediction method, amino acid hydropathy scale used, and the prediction window size. We analyzed altogether 56 scales which have been presented for amino acid hydropathy. Since nearly all prediction methods in literature use the sliding window technique, we also used that approach. First we optimized the length of the window i.e. the number of residues for which the hydropathy parameter is calculated at a time.

Literature contains numerous hydropathy scales, many of which were reviewed and normalized [12] to study the performance of the scales in detecting amphipathic helices. Altogether 39 of the scales were complete i.e. parameters were given for all residue types. In some of the old scales parameters were incomplete. Some new scales have been presented. We developed a method to make and present hydropathy predictions with 17 additional scales [13]. To make the scales comparable they were normalized.

2.1. Optimization of the prediction window

The effect of the window size was tested by calculating the correlation for all the scales by using window sizes of odd numbers between 5 and 25. Odd numbers were used because the resulting value can be given to the middlemost residue within the window. The results are for the 2441 protein structures investigated.

The window size of 9 consecutive residues was the best for all the scales. The length of the window has a great effect on the prediction accuracy. For the best scales the accuracy varied from -0.12 to -0.26 just by changing the window size. In addition, the order of accuracy of the windows is the same for all the scales indicating that this is an intrinsic property of the hydropathy sliding window method. In all subsequent analyses the window size 9 was used.

We have previously studied the accuracy of protein flexibility predictions [14]. Flexibility indicates how mobile a residue is within a protein. The tested flexibility predictions had correlations in the order of 0.33, whereas the best hydropathy correlations are approximately -0.26 and the worst about -0.11 (Table 1). The correlations are negative because hydrophilic residues have negative values due to our normalization, which was used to organize the scales. The five best scales are those of Nioii [15], Mijer [16], Ponnu [17], Nneig [12] and Guym [18]. Three of these scales are based on statistical amino acid distribution analysis (Nioii, Mijer and Nneig) while one is a combined experimental and statistical scale (Ponnu), and one an average scale (Guym). Thus, different types of scales can provide similarly reliable predictions. However, among the best scales were no experimental scales, the first being in 16th place (Abodr [19]). Since differences in the overall accuracy are small at the beginning of the list, Abodr is not significantly worse than the top scales. When looking at the methods giving the poorest predictions (Jones [20], Ponma [21], Zimmr [22], IHF1 [23] and Urry [24]), four of these scales are experimental, Ponma being the only statistical scale. Thus, experimental data alone is not sufficient for the development of accurate hydropathy prediction methods. The least accurate methods are significantly worse than the best scales. The hydropathy characteristics are clearly structure and context dependent, so therefore amino acids, fragments or peptide based experimental scales cannot reliably explain the hydropathy of residues in folded three dimensional protein structures.

The overall correlation is not very high (Table 1), however since these methods are primarily used to search for the most hydrophilic or hydrophobic regions the applicability is higher than the plain numbers indicate. The reason for the low correlation is that amino acids within e.g. exposed structures are not of one single type. Also protein folding affects the environments e.g. in β-strands every second residue is pointing in the opposite direction. Amphipathy is relatively common in secondary structural elements. Predictions based on linear sequence do not provide very accurate predictions for each individual residue, but observable trends. Predictions of hydrophilic regions have been used for example to raise antibodies. The shapes of the surface accessibility and hydropathy predictions in Figure 2 are relatively

similar despite the low correlation. These results clearly point that hydropathy predictions can still be useful for some qualitative purposes, although one has to be diligent when choosing the method and applying the obtained results.

2.2. Prediction accuracy of amino acids

There are several groups of amino acid residues with similar or related characteristics; such as positively and negatively charged residues, aliphatic amino acids etc. Amino acids have different distribution for accessible and buried residues ([25]). However, practically all residues appear both at buried and exposed sites. There are several other features that lead to different distributions of residues in e.g. secondary structures and 3D folds.

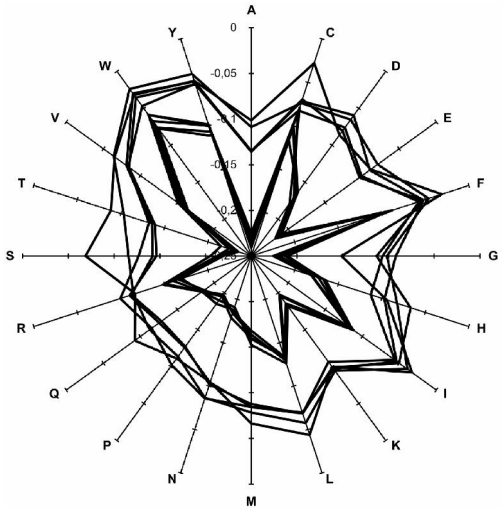


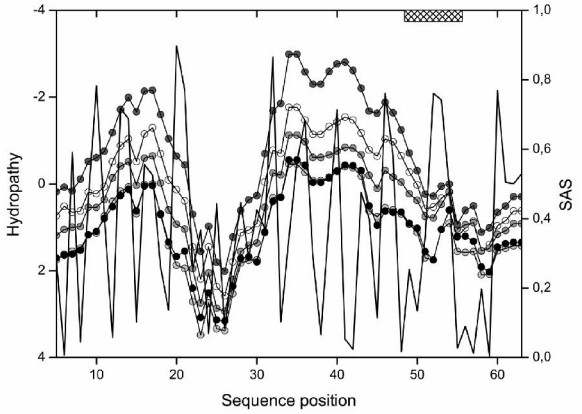
Figure 1. Prediction accuracy of individual amino acids on the five most (Nioii, Mijer, Ponnu, Nneig, Guym) and least (Jones, PONMA, Zimmr, IFH1, Urry) accurate scales.

The hydropathy predictions were best for residues A, E, G, K, P, S and T (≤-0.20) (Figure 1). The correlations are in the order of -0.3, whereas the worst predictions for C, I, L, F, W and Y are merely around -0.1. Among the well predicted residues are A and G, the most flexible residues, and P, which has restricted main chain torsion angles. Three of the poorly predicted residues are bulky aromatic amino acids and two of them are aliphatic (I, L). H also has relatively poor overall prediction accuracy. The results are very similar for all the methods i.e. the same residues have poor or good predictions independent of the scale.

The scales were normalized so that glycine has value 0 and hydrophilic amino acids have negative values. Values in some scales were so skewed that glycine was not given value 0 [12]. The differences in the values between the scales are very high for some residues, even so that the scales do not agree whether a residue is hydrophilic or hydrophobic. The values for amino acids in the scales for the best methods are close to each other, although they are based on different principles.

2.3. Correlation of hydropathy predictions to linear epitope regions

One widely used biological application of hydropathy scales is to predict the location of antigenic regions, i.e. epitopes. Based on these predictions, antibodies have been produced against synthetic peptides and used to bind to full length proteins. Epitopes can either be conformational or linear. Hydropathy predictions cannot detect conformational epitopes because they are formed by regions that fold together from different parts of the polypeptide chain. Linear epitopes are formed by consecutive amino acids in exposed regions on the protein surface. Hydropathy predictions should be able to detect such regions.



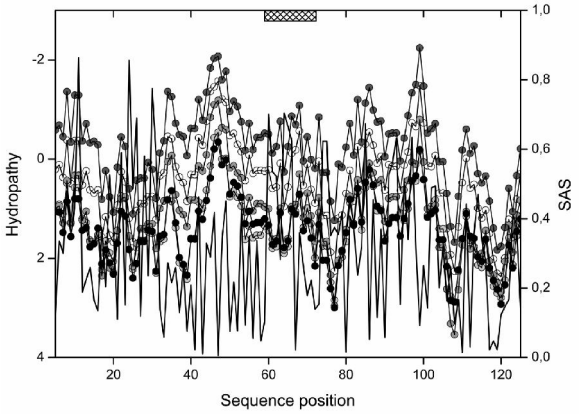


Figure 2. **A** Correlation of hydropathy predictions and SAS volumes for the C-terminal zinc binding domain of the HPV16 E6 oncoprotein (PDB code 2fk4). **B** Correlation of hydropathy predictions and SAS volumes for the major house dust mite allergen Der p 2 (1a9v). The hydropathy predictions with the five most reliable hydropathy scales are presented Nioii (○), Mijer (○ light gray), Ponnu (○ gray), Nneig (○ dark gray) and Guym (•) along with the surface accessible volume of amino acids (black). The locations of linear epitopes are indicated with boxes on top.

Table 1. Correlation coefficients for hydropathy predictions with all proteins.

Scale	r	Ref	Scale	r	Ref
Nioii	-0,258	15	AMP07	-0,233	40
Mijer	-0,258	16	Faupl	-0,231	41
Ponnu	-0,258	17	MPH89	-0,228	40
Nneig	-0,258	12	Choth	-0,227	42
Guym	-0,257	18	Miller	-0,221	43
Meiro	-0,256	26	Hoppw	-0,220	44
Totls	-0,256	12	Prift	-0,213	12
Sweet	-0,254	27	Eisen	-0,210	45
PKPON	-0,254	28	Olsen	-0,210	46
Rosef	-0,252	29	Fromm	-0,207	47
Cassi	-0,252	30	Clogp	-0,199	48
Holbro	-0,251	31	Levit	-0,198	49
Prils	-0,251	12	Rosem	-0,198	50
Sweig	-0,251	12	Jadlg	-0,197	18
Vhebl	-0,250	32	Chdlg	-0,190	18
Abodr	-0,249	19	Engel	-0,190	51
Wersc	-0,248	33	Zwitter	-0,188	40
Krigk	-0,246	34	Meek	-0,185	52
Wsdlg	-0,246	18	Eimcl	-0,183	53
Janmi	-0,246	35	Kridg	-0,181	52
Totft	-0,246	12	Kuntz	-0,181	54
Alft	-0,245	12	1HF0	-0,161	23
Altls	-0,244	12	1HF0.5	-0,156	23
KytDo	-0,241	36	Jones	-0,150	20
Janin	-0,237	37	PONMA	-0,144	21
Guy	-0,237	18	Zimmr	-0,143	22
Buldg	-0,234	38	1HF1	-0,142	23
PARKER	-0,233	39	Urry	-0,111	24

We tested the usability of the best hydropathy scales by predicting epitopes for a number of proteins for which the structure and location of linear epitopes is known. The data was taken from the Immune Epitope Database and Analysis Resource (IEDB) [55]. Figure 2 indicates that the B-cell epitopes are indeed predicted to have hydrophilic peaks. Instead of long hydrophilic runs, the predictions have peaks within the epitope regions. The predictions are not necessarily highest for epitopes. The house dust mite allergen Der p2 structure (Figure 2B) indicates that epitope regions can be relatively reliably predicted even when the overall correlation is low (-0.03 to -0.09). For the other proteins in the figures, the correlations range from -0.14 to -0.44. It is difficult to elucidate which ones of the hydrophilic peaks represent epitopes. It could therefore be necessary to synthetize several peptides and produce several antibodies, which is costly. Combination with other epitope prediction methods could possibly increase the success rate. Recently, in addition to amino acid propensity scale based methods (e.g. 12, 13) some more advanced machine learning methods have been released based on e.g. neural networks [56], hidden Markov model [57], and Gibbs sampling [58].

3. Conclusions

We investigated the prediction accuracy of numerous hydropathy scales. A window size of 9 proved to be the best for all the scales. Window size has a strong effect on the accuracy. The correlations for the best scales are in the order of -0.26. The overall correlation is very low. The sliding window technique smoothes the predictions over several residues. Therefore the method cannot predict correctly alternating exposed/buried sequence stretches for example.

Results for different amino acids vary greatly within each scale, but are more consistent between the scales. Hydropathy scales have been determined based on many principles. Experimental scales measuring the partition between phases are on average the poorest. This implies that folded protein structures have many additional features not measured when investigating just fragments, amino acids or short peptides.

Previously correlation of the hydropathy scales [59] was investigated for five scales. They obtained correlation coefficients in the range of 0.40 to 0.45. The difference in correlations to those measured by us is due the method they applied. They used an average sequence for each position build based on multiple sequence alignment in addition the average accessibility was calculated from a family of proteins. This is a good approach when working with a family of sequences and several structures are available. However, hydropathy predictions are mostly used when this kind of data is unavailable and the normal sliding window approach is the solution. In protein families with several known structures no need for predictions exists.

The accuracy of linear epitope predictions have been assessed for 484 amino acid property scales including several hydropathy scales [60]. This analysis concentrated on peak values and predictions of epitopes for 50 proteins. The correlation was very poor even for the best scales (in the order of 0.15).

In conclusion, hydropathy scales are widely used without any idea about how reliable the results are. We performed the first large scale analysis to correlate hydropathy scales with surface exposure of residues in 3-D structures. Hydropathy predictions have numerous applications and are widely used by biologists. For many purposes trends in hydropathy are more important than actual values. Despite low overall correlation, hydropathy predictions can still be used for some applications but only when qualitative analysis is sufficient.

4. Methods

We studied the accuracy of the hydropathy scales by calculating the Pearson correlations between predicted amino acid hydropathy and calculated accessible surface areas for protein 3-D structures obtained from the PDB [61]. The proteins were taken from the 25% list of PDBselect [62]. These structures are not more than 25% identical at sequence level. 2441 structures were used in the analysis.

The hydropathy predictions were calculated by using the sliding window averaging technique. The hydrophobicity scale assigns a hydropathy value to each amino acid and a mean hydropathy value was calculated within the window. The mean value was given for the middlemost amino acid in the window and repeated in steps of one amino acid for the whole sequence. Window lengths of odd numbers between 5 and 25 amino acids were analyzed. The hydropathy scales were normalized so that the values vary between -10 and 10. Glycine was assigned a zero, hydrophobic amino acids have positive values and hydrophilic negative values.

The amino acid sequences and accessibility values were derived with STRIDE [63], which uses both hydrogen bond energy and main chain dihedral angles to recognize secondary structural elements. Solvent accessible surface (SAS) values for each amino acid calculated with STRIDE were divided by the maximum value of the amino acid. The amino acid maximum accessibility values were determined from an extended conformation where residues were surrounded by glycines on both sides.

For correlations we used Pearson Product Moment Correlation

$$r = \frac{\sum x_i y_i - \sum x_i \sum y_i / N}{((\sum x_i^2 - (\sum x_i)^2/N) * (\sum y_i^2 - (\sum y_i)^2/N))},$$

where r is Pearson's correlation coefficient, xi is the mean hydropathy value within a window, yi is the maximum value divided SAS value corresponding xi, and N denotes the number of amino acids. The analysis was extended by calculating Pearson correlations between accessibility values and hydropathy values for the 56 scales for the 20 amino acids.

5. Acknowledgements

We want to thank the Medical Research Fund of Tampere University Hospital and the Department of Information Technology in the University of Turku for financial support and Kathryn Rannikko for language correction.

6. References

1. Dill K. Dominant forces in protein folding. *Biochemistry* 1990, 29, pp. 133-7155.

2. Vitkup D, Sander C, Church G. The amino-acid mutational spectrum of human genetic disease. *Genome Biol* 2003, 4. R72.
3. Eriksson A, Baase W, Zhang X, Heinz D, Blaber M, Baldwin E, Matthews B. Response of a protein structure to cavity-creating mutations and its relation to the hydrophobic effect. *Science* 1992, 255, pp. 178-183.
4. Lolkema J, Slotboom D. Hydropathy profile alignment. a tool to search for structural homologues of membrane proteins. *FEMS Microbiol. Rev* 1998, 22, pp. 305-322.
5. Clements JD, Martin RE. Identification of novel membrane proteins by searching for patterns in hydropathy profiles. *Eur. J. Biochem* 2002, 269, pp. 2101-2107.
6. Schwarz R, King J. Frequencies of hydrophobic and hydrophilic runs and alternations in proteins of known structure. *Protein Sci* 2006, 15, pp.102-112.
7. Baranyi L, Campbell W, Ohshima K, Fujimoto S, Boros M, Okada H. The antisense homology box. a new motif within proteins that encodes biologically active peptides. *Nat. Med* 1995, 1, pp. 894-901.
8. Berchanski A, Shapira B, Eisenstein M. Hydrophobic complementarity in protein-protein docking. *Proteins* 2004, 56, pp. 130-142.
9. Viklund H, Elofsson A. Best α-helical transmembrane protein topology predictions are achieved using hidden Markov models and evolutionary information. *Protein Sci* 2004, 13, pp. 1908-1917.
10. Bagos P, Liakopoulos T, Hamodrakas S. Evaluation of methods for predicting the topology of β-barrel outer membrane proteins and a consensus prediction method. *BMC Bioinformatics* 2005, 6, pp. 7.
11. Palliser C, Parry D. Quantitative comparison of the ability of hydropathy scales to recognize surface β-strands in proteins. *Proteins* 2001, 42, pp. 243-255.
12. Cornette J, Cease K, Margalit H, Spouge J, Berzofsky J, DeLisi C. Hydrophobicity scales and computational techniques for detecting amphipathic structures in proteins. *J. Mol. Biol* 1987, 195, pp. 659-685.
13. Vihinen, Torkkila E. HYDRO. a program for protein hydropathy predictions. *Comput. Methods Programs Biomed* 1993, 41, pp. 121-129.
14. Vihinen M, Torkkila E, Riikonen P. Accuracy of protein flexibility predictions. *Proteins* 1994, 19, pp. 141-149.
15. Nishikawa K, Ooi T. Prediction of the surface-interior diagram of globular proteins by an empirical method. *Int. J. Pept. Protein Res* 1980, 16, pp. 19-32.
16. Miyazawa S, Jernigan R. Estimation of effective interresidue contact energies from protein crystal structures. Quasi-chemical approximation. *Macromolecules* 1985, 18, pp. 534-552.
17. Ponnuswamy P, Prabhakaran M, Manavalan P. Hydrophobic packing and spatial arrangement of amino acid residues in globular proteins. *Biochim. Biophys. Acta* 1980, 623, pp. 301-316.
18. Guy H. Amino acid side-chain partition energies and distribution of residues in soluble proteins. *Biophys. J* 1985, 47, pp. 61-70.
19. Aboderin A. An empirical hydrophobicity scale for alpha-amino-acids and some of its applications. *Int. J. Biochem* 1971, 2, pp. 537-544.
20. Jones D. Amino acid properties and side-chain orientation in proteins. a cross correlation approach. *J. Theor. Biol* 1975, 50, pp. 2577-2637.
21. Manavalan P, Ponnuswamy P. 1978. Hydrophobic character of amino acid residues in globular proteins. *Nature* 275. 673-674.
22. Zimmerman J, Eliezer N, Simha R. The characterization of amino acid sequences in proteins by statistical methods. *J. Theor. Biol* 1968, 21, pp. 170-201.
23. Jacobs R, White S. The nature of the hydrophobic binding of small peptides at the bilayer interface. implications for the insertion of transbilayer helices. *Biochemistry* 1989, 28, pp. 3421-3437.
24. Urry D, Gowda C, Parker T, Luan C, Reid M, Harris C, Pattanaik A, Harris, R. Hydrophobicity scale for proteins based on inverse temperature transitions. *Biopolymers* 1992, 32, pp. 1243-1250.
25. Lawrence C, Auger I, Mannella C. Distribution of accessible surfaces of amino acids in globular proteins. *Proteins* 1987, 2, pp. 153-61.
26. Meirovitch H, Rackovsky S, Scheraga H. Empirical studies of hydrophobicity. 1. Effect of protein size on the hydrophobic behavior of amino acids. *Macromolecules* 1980, 13, pp. 1398-1405.
27. Sweet R, Eisenberg D. Correlation of sequence hydrophobicities measures similarity in three-dimensional protein structure. *J. Mol. Biol* 1983, 171, pp. 479-488.
28. Ponnuswamy P. Hydrophobic characteristics of folded proteins. *Prog. Biophys. Mol. Biol* 1993, 59, pp. 57-103.
29. Rose G, Geselowitz A, Lesser G, Lee R, Zehfus M. Hydrophobicity of amino acid residues in globular proteins. *Science* 1985, 229, pp. 834-838.

30. Casari G, Sippl M. Structure-derived hydrophobic potential. Hydrophobic potential derived from X-ray structures of globular proteins is able to identify native folds. *J. Mol. Biol* 1992, 224, pp. 725-732.
31. Holbrook S, Muskal S, Kim SH. Predicting surface exposure of amino acids from protein sequence. *Protein Eng* 1990, 3, pp. 659-665.
32. von Heijne G, Blomberg C. Transmembrane translocation of proteins. The direct transfer model. *Eur. J. Biochem* 1979, 97, pp. 175-181.
33. Wertz D, Shegara H. Influence of water on protein structure. An analysis of the preferences of amino acid residues for the inside or outside and for specific conformations in a protein molecule *Macromolecules* 1978, 11, pp. 9-15.
34. Krigbaum W, Komoriya A. Local interactions as a structure determinant for protein molecules. II. *Biochim. Biophys. Acta* 1979, 576, pp. 204-248.
35. Janin J, Miller S, Chothia C. Surface, subunit interfaces and interior of oligomeric proteins. *J. Mol. Biol* 1988, 204, pp. 155-164.
36. Kyte J, Doolittle R. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol* 1982, 157, pp. 105-132.
37. Janin J: Surface and inside volumes in globular proteins. *Nature* 1979, 277, pp. 491-492.
38. Bull H, Breese K. Surface tension of amino acid solutions. a hydrophobicity scale of the amino acid residues. *Arch. Biochem. Biophys* 1974, 161, pp. 665-670.
39. Parker J, Guo D, Hodges R. New hydrophilicity scale derived from high-performance liquid chromatography peptide retention data. correlation of predicted surface residues with antigenicity and X-ray-derived accessible sites. *Biochemistry* 1986, 25, pp. 5425-5432.
40. Degli Esposti M, Crimi M, Venturoli G. A critical evaluation of the hydropathy profile of membrane proteins. *Eur J Biochem* 1990, 190, pp. 207-219.
41. Fauchère J, Pliška V. Hydrophobic parameters π of amino-acid side chains from the partitioning of N-acetyl-amino-acid amides. *Eur. J. Med. Chem* 1983, 18, pp. 369-52.
42. Chothia C. The nature of the accessible and buried surfaces in proteins. *J. Mol. Biol* 1976, 105, pp. 1-12.
43. Miller S, Janin J, Lesk A, Chothia C. Interior and surface of monomeric proteins. *J. Mol. Biol* 1987, 196, pp. 641-656.
44. Hopp T, Woods K. Prediction of protein antigenic determinants from amino acid sequences. *Proc. Natl. Acad. Sci* 1981, 78, pp. 3824-3828.
45. Eisenberg D, Weiss R, Terwilliger T, Wilcox W. Hydrophobic moments and protein structure. *Faraday Symp. Chem. Soc* 1982b, 17, pp. 109-120.
46. Olsen K. Internal residue criteria for predicting three-dimensional protein structures. *Biochem. Biophys. Acta* 1980, 622, pp. 259-267.
47. Frömmel C. The apolar surface area of amino acids and its empirical correlation with hydrophobic free energy. *J Theor Biol* 111, pp. 247-260.
48. Abraham D, Leo A. Extension of the fragment method to calculate amino acid zwitterion and side chain partition coefficients. *Proteins* 1987, 2, pp. 130-152.
49. Levitt M. A simplified representation of protein conformations for rapid simulation of protein folding. *J. Mol. Biol* 1976, 104, pp. 59-107.
50. Roseman M.A. Hydrophilicity of polar amino acid side-chains is markedly reduced by flanking peptide bonds. *J. Mol. Biol* 1988, 200, pp. 513-522.
51. Engelman D, Steitz T, Goldman A. Identifying nonpolar transbilayer helices in amino acid sequences of membrane proteins. *Annu. Rev. Biophys. Biophys. Chem* 1986, 15, pp. 321-353.
52. Meek J. Prediction of peptide retention times in high-pressure liquid chromatography on the basis of amino acid composition. *Proc. Natl. Acad. Sci* 1980, 77, pp. 1632-1636.57.
53. Eisenberg D, MacLachan A. Solvation energy in protein folding and binding. *Nature* 1986, 319, pp. 199-203.
54. Kuntz I. Hydration of macromolecules. IV. Polypeptide conformation in frozen solutions. *J. Amer. Chem.* Soc 1971, 93, pp. 516-518.
55. Peters B, Sidney J, Bourne P, Bui H, Buus S, Doh G, Fleri W, Kronenberg M, Kubo R, Lund, O, Nemazee D, Ponomarenko J, Sathiamurthy M, Schoenberger S, Stewart S, Surko P, Way, S, Wilson S, Sette A. The immune epitope database and analysis resource. from vision to blueprint. *PLoS Biol* 2005, 3. e91.
56. Saha S, Raghava G.P. Prediction of continuous B-cell epitopes in an antigen using recurrent neural network. Proteins 2006, 1, 65(1), pp. 40-8.
57. Larsen J.E, Lund O, Nielsen M. Improved method for predicting linear B-cell epitopes. *Immunome Res.* 2006, 24, 2.2.
58. Nielsen M, Lundegaard C, Worning P, et al.. Improved prediction of MHC class I and class II epitopes using a novel Gibbs sampling approach. *Bioinformatics* 2004, 20(9), pp. 1388-97.
59. Moelbert S, Emberly E, Tang C. Correlation between sequence hydrophobicity and surface

exposure pattern of database proteins. *Protein Sci* 13, pp. 752-762.
60. Blythe M, Flower D. Benchmarking B cell epitope prediction. Underperformace of existing methods *Prot. Sci*. 2005, 14, pp. 246-248.
61. Berman H, Henrick K, Nakamura H. Announcing the worldwide Protein Data Bank. *Nature Struct. Biol* 2003, 10, pp. 980.
62. Hobohm U, Scharf M, Schneider, Sander C. Selection of representative protein data sets. *Protein Sci* 1992, 1, pp. 409-417.
63. Frishman D, Argos P. Knowledge-based secondary structure assignment. *Proteins* 1995, 23, pp. 566-579.