Progress in Biophysics and Molecular Biology xxx (2017) 1–10

Contents lists available at ScienceDirect

Progress in Biophysics and Molecular Biology

journal homepage: www.elsevier.com/locate/pbiomolbio

Depth dependent amino acid substitution matrices and their use in predicting deleterious mutations

Nida Farheen [a, 1], Neeladri Sen [a, 1], Sanjana Nair [a], Kuan Pern Tan [b, c], M.S. Madhusudhan [a, b, *]

[a] Indian Institute of Science Education and Research, Pune 411008, India
[b] Bioinformatics Institute, 30 Biopolis Street, #07-01, Matrix, Singapore 138671
[c] School of Computer Engineering, Nanyang Technological University, Singapore 639798

ARTICLE INFO

Article history:
Received 13 September 2016
Received in revised form
6 January 2017
Accepted 7 February 2017
Available online xxx

Keywords:
Substitution matrix
Alignment
Depth
Deleterious mutation

ABSTRACT

The 20 naturally occurring amino acids have different environmental preferences of where they are likely to occur in protein structures. Environments in a protein can be classified by their proximity to solvent by the residue depth measure. Since the frequencies of amino acids are different at various depth levels, the substitution frequencies should vary according to depth. To quantify these substitution frequencies, we built depth dependent substitution matrices. The dataset used for creation of the matrices consisted of 3696 high quality, non redundant pairwise protein structural alignments. One of the applications of these matrices is to predict the tolerance of mutations in different protein environments. Using these substitution scores the prediction of deleterious mutations was done on 3500 mutations in T4 lysozyme and CcdB. The accuracy of the technique in terms of the Matthews Correlation Coefficient (MCC) is 0.48 on the CcdB testing set, while the best of the other tested methods has an MCC of 0.40. Further developments in these substitution matrices could help in improving structure-sequence alignment for protein 3D structure modeling.

Contents

* Corresponding author. Indian Institute of Science Education and Research, Pune 411008, India.
E-mail address: madhusudhan@iiserpune.ac.in (M.S. Madhusudhan).
[1] Equal contribution.

http://dx.doi.org/10.1016/j.pbiomolbio.2017.02.004
0079-6107/

1. Introduction

The 3D structure of a protein is key to determining its function or biological role. The primary sequence of a protein folds into a particular 3D shape, given a particular set of conditions (Anfinsen, 1973). The number of shapes that proteins fold into are limited, and by various estimates, is of the order of 1000 (Chothia, 1992). It is believed that the native fold of a protein is its minimum energy conformation (Wolynes et al., 1995). This conformation is solely dependent on its amino acid sequence. Any changes to the amino acid chain could result in a perturbation of this 3D structure.

With respect to mutations of amino acids in proteins, one of the key questions to answer, would be to determine if the mutations could change the conformation of the protein sufficiently to affect function. Note that the function of a protein could also be affected by mutations that need not necessarily change its 3D shape. For the purposes of this study we are only interested in those mutations that affect the stability of the 3D structure of the protein. Our motivation arises form that fact that ~80% of the Mendelian-disease-associated single mutations are a consequence of protein destabilization (Wang and Moult, 2001).

The effects of a single point mutation in a protein sequence are most acutely felt by its immediate spatial neighbours. In essence, every single amino acid in a protein is embedded in its own characteristic microenvironment. Traditionally, solvent accessible surface area (SASA) (Lee and Richards, 1971) was one of the ways in which these microenvironments were categorized. SASA values were classified into levels such as buried, intermediate and exposed. Residues in the hydrophobic core of a globular protein were typically buried while the polar residues that constituted the periphery of the protein were exposed. This classification however is rather coarse (Fig. 1a) and does not stratify the interior of the protein adequately. A more concise description of the residue environment is provided by the depth measure (Chakravarty and Varadarajan, 1999). Residue (or atom) depth is defined as the distance of a residue (or atom) to the closest molecule of bulk solvent. This definition offers a more stratified description of the protein interior (Fig. 1b).

That residue depth is an apt descriptor of protein microenvironments is further evidenced from the many uses of depth. Residue depth correlates better with hydrogen-deuterium exchange data than SASA (Chakravarty and Varadarajan, 1999). It is also a vital feature in the detection of post translational modification sites (Pintar et al., 2003a, 2003b). In conjunction with SASA, depth has been used to predict small molecule ligand binding sites and cavities in proteins (Tan et al., 2013, 2011). Combining depth with SASA, electrostatic and hydrogen bonding interactions has been shown to effectively predict the pK_a of ionizible groups in proteins (Tan et al., 2013). Residue depth has been efficiently combined with hydrophobicity and hydrophobic moment derived from the primary sequence of the protein to predict temperature sensitive mutations (Tan et al., 2014). In combination with evolutionary sequence profiles and SASA, depth could be used to recognize native protein folds (Liu et al., 2007; Zhou and Zhou, 2005). In each of the applications mentioned above the key aspect has been the ability of depth to describe the immediate neighbourhood of amino acid residues. In this study we are going to utilize this feature of depth to determine how the immediate neighbourhood of an amino acid is affected on mutation.

An observation that is crucial to our study is that the amino acid abundance at different depth levels is markedly different (Fig. 2). The depth preferences of some of the amino acids could be categorized based on the nature of their side chains. The polar amino acids (N, Q, H, K, R, D, E) show a sharp decline in their abundance with increase in depth. The hydrophobic amino acids (V, I, L, M, F)

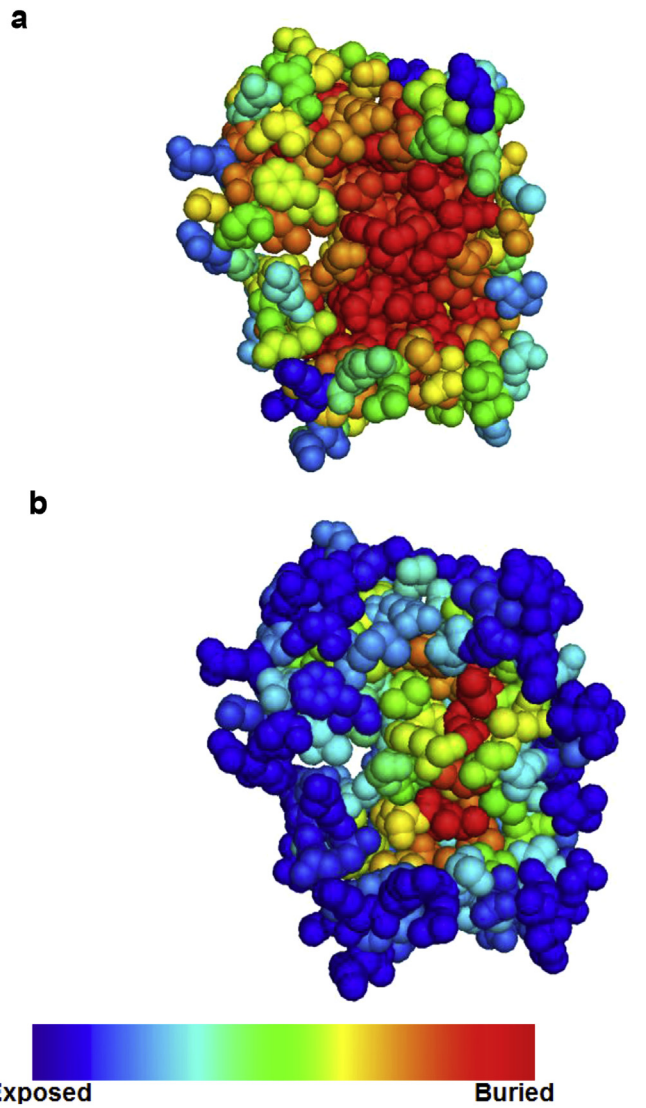

Fig. 1. A cross section of a protein (human dihydrofolate reductase, PDB id 1MVT) stratifying microenvironments by (a) SASA (b) Depth. All atoms of the protein are rendered in sphere representation and are coloured according to SASA and depth using PyMol (DeLano, 2002).

have an increase in abundance with increase in depth. The amino acids S, T and G also behave like the polar amino acids only that the decrease in abundance is not sharp. The amino acids A, Y and W have their maximum abundance in an environment that is neither deep nor shallow. Cysteines, though considered polar by some studies, show the same behavior as non-polar residues while Prolines, which is sometimes considered apolar, displays the same tendency as polar residues. It is clear that with stratification by depth, relative abundances of amino acids vary. We use this fact to compute the likelihoods of amino acid substitutions. Note that these trends are best observed by parameterizing protein environment using Depth as opposed to SASA (Chakravarty and Varadarajan, 1999).

Computations of substitution likelihoods have been well documented (Dayhoff and Schwartz, 1978; Henikoff and Henikoff, 1992) and widely used from aligning two sequences to one another to detecting homologous sequences (Altschul et al., 1990; Chenna et al., 2003). The traditional substitution likelihoods bundled into the so called substitution matrices, such as PAM and BLOSUM, are

Please cite this article in press as: Farheen, N., et al., Depth dependent amino acid substitution matrices and their use in predicting deleterious mutations, Progress in Biophysics and Molecular Biology (2017), http://dx.doi.org/10.1016/j.pbiomolbio.2017.02.004

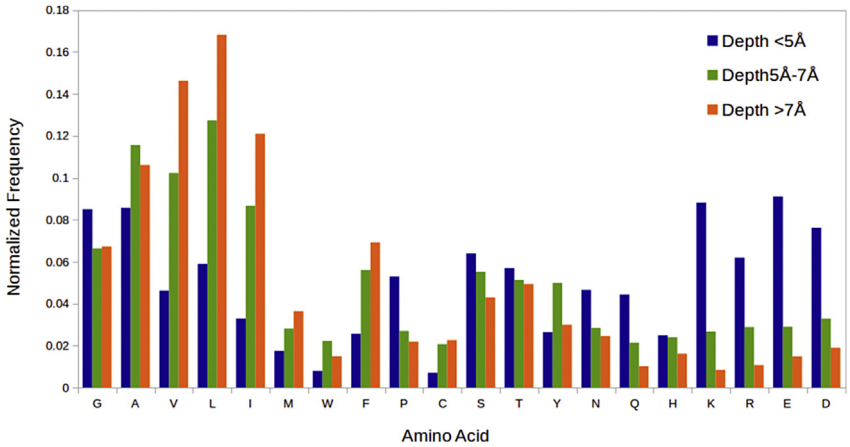

Fig. 2. Histograms of the relative abundance of amino acids at different depth levels <5 Å (blue), 5–8 Å (green) and>8 Å (orange). Normalization of the abundance was done depth wise.

however devoid of any context. In fact, amino acid substitutions involving a pair of residues is averaged over several different environmental and fold contexts. In the light of secondary structure prediction and 3D structure modeling/evaluation exercises, amino acid environments have to be described accurately. This implies that an amino acid substitution table that considered not just the likelihoods of pairwise substitutions but also the environmental context and/or protein categories would be best suited for the purpose (Abascal et al., 2007; Adachi and Hasegawa, 1996; Arvestad, 2006; Braberg et al., 2012; Dimmic et al., 2002; Goldman et al., 1998; Johnson et al., 1993; Jones et al., 1994; Koshi and Goldstein, 1995; Lartillot and Philippe, 2004; Lüthy et al., 1991; Madhusudhan et al., 2009; Mehta et al., 1995; Overington et al., 1990; Rice and Eisenberg, 1997; Shi et al., 2001; Thorne et al., 1996; Wako and Blundell, 1994). As depth is a concise measure of amino acid environment, we developed depth dependent substitution matrices that can capture the substitution likelihoods in different environments. In this study we have categorized amino acids into 3 distinct environments – residues that lie in depth ranges <5 Å, between 5 and 8 Å and >8 Å. As described earlier, the relative abundance of the residues in these depth environments are different and hence it is likely that their substitution rates would also differ in the different contexts. A symmetric substitution matrix was computed for each of the depth environments considering a log-odds ratio of observed over expected frequencies.

The efficacy of the matrices was tested by using it to predict the destabilizing effects of single point mutations in protein sequences. Other computational methods that address this question use a combination of sequence and structural information to deduce the effect of the mutation. The approaches for predicting the mutational stability could be divided into sequence-based and structure-based methods. Sequence based methods such as SIFT (Ng and Henikoff, 2003), Polyphen (Adzhubei et al., 2010) and SuSPect (Yates et al., 2014) rely on multiple sequence alignments of proteins to extract substitution trends from sequence profiles. Polyphen and SuSPect also utilize structure features. SuSPect incorporates the extraction of information from protein domains, PSSM, protein-protein network interactions, position-specific known mutants and is one of the methods compared to in this study. Most structure-based methods are based on machine learning that fit a non-linear function to experimental data. We have compared our method (FADHM) to several such methods including I-Mutant (Capriotti et al., 2005), Automute (Masso and Vaisman, 2010), mCSM (Pires et al., 2014a), SDM(Worth et al., 2011) and DUET (Pires et al., 2014b). I-Mutant incorporates pH, temperature and mutation type as features in its support vector machine. Automute is based on a multi-body statistical potential that combines energy-based and machine learning approaches. mCSM (Pires et al., 2014a) uses a graph metric to summarize physiochemical interactions within a cutoff distance and train them with a Gaussian process regression model. SDM(Worth et al., 2011) is a statistical method that builds an environment dependent substitution matrix. DUET (Pires et al., 2014b) is a meta-algorithm combining mCSM and SDM.

Predictions made by the depth dependent substitution matrices were benchmarked using saturation mutagenesis data available for T4 Lysozyme (Rennell et al., 1991) and *E. coli* Controller of Cell Division or Death B (CcdB) protein (Adkar et al., 2012; Tripathi et al., 2016). The accuracy of our predictions were compared to those made by other methods described above.

2. Methods

2.1. Computation of residue depth

Depth is a concise descriptor of amino acid residue environment (Chakravarty and Varadarajan, 1999; Pintar et al., 2003a, 2003b; Tan et al., 2013, 2011). It is defined as the average distance of the atoms of the residue to their nearest bulk solvent. In this study, residue depth was computed by previously described methods (Tan et al., 2013, 2011), using default parameters. Here, we have only considered protein structures that had only a few or no missing residues (see section 2.2). Missing residues could alter the distance to the closest molecule of bulk solvent and hence affect depth values.

2.2. Pairwise alignments for matrix creation

1607 structures were culled from the protein data bank (PDB) (Berman et al., 2000)using PISCES(Wang and Dunbrack, 2003) and home grown scripts such that their a) sequences were non-redundant at 30% sequence identity, b) resolution was <3 Å with R-factor < 0.3 and c) structures were missing fewer than 6 contiguous residues. Missing stretches were modeled using the loop modeling (Fiser et al., 2000) module of MODELLER (Sali and Blundell, 1993). Structures that had more than 6 missing residues were discarded, as errors in loop modeling could be significant and introduce errors in depth measurements.

Please cite this article in press as: Farheen, N., et al., Depth dependent amino acid substitution matrices and their use in predicting deleterious mutations, Progress in Biophysics and Molecular Biology (2017), http://dx.doi.org/10.1016/j.pbiomolbio.2017.02.004

BLAST (Altschul et al., 1990) was used to identify the homologues of these 1607 proteins from the PDB. From this, 1426 homologues of 947 structures (from the initial 1607) were chosen such that the e-values were less than 0.001 and pair-wise sequence identities were less than 30%. From these 2383 (1426 + 947) structures, 3696 pair wise structure-structure alignments were constructed using SALIGN (Braberg et al., 2012) such that the SALIGN quality score was ≥ 85% and the length difference between the 2 aligned proteins was <35 residues. These alignments gave us 800,558 residue substitutions.

2.3. Creation of depth dependent substitution matrices

Multiple substitution matrices were created from the pair-wise structure-structure alignments. All matrix values, $S^d_{i,j}$, were ratios of observed over expected residue substitution likelihoods and were computed using similar formulae used in BLOSUM(Henikoff and Henikoff, 1992).

$$S^d_{i,j} = 2 \cdot log_2 \left(\frac{q^d_{i,j}}{e^d_{i,j}} \right) \quad (1)$$

where i and j are the residues that are being substituted to one another and d being the depth of the residue i. Note that in these matrices, substitution of $i$to$j$ is considered equivalent to that of $j$to$i$. $q^d_{i,j}$ is the observed substitution probability and $e^d_{i,j}$ is the expected probability. The matrix values are scaled by a factor of $2 \cdot log_2$, similar to the BLOSUM62 matrix.

The observed probability is computed as

$$q^d_{i,j} = \frac{f^d_{i,j}}{\sum_{i=1}^{20} \sum_{j=1}^{20} f^d_{i,j}} \quad (2)$$

where $f^d_{i,j}$ is the number of substitutions of residue i to j (and vice versa) at depth range d. The denominator of equation (2) is the total number of observed residue substitutions.

The expected probability of residue substitution at the different depth ranges is given by

$$e^d_{i,j} = \begin{cases} p^d_i * p^d_j \ when \ i = j \\ 2 \cdot p^d_i * p^d_j \ when \ i \neq j \end{cases} \quad (3)$$

where,p^d_i is the probability of residue i at depth range d and is given by

$$p^d_i = q^d_{i,j} + \sum_{i \neq j} \frac{q^d_{i,j}}{2} \quad (4)$$

2.4. Database of single point mutants

Depth dependent substitution matrices were used to predict the effect of single point mutations in proteins. The predictions were trained on 1966 mutations of T4 Lysozyme (Rennell et al., 1991), where 163 of the 164 amino acids of the protein were mutated to one of 13 different amino acids (A, C, E, F, G, H, K, L, P, Q, R, S, and T) after removal of the key catalytic residues (D10, E11, R145, and R148). The prediction training was done using a grid search over substitution values (searched in a range of −3.00 to 1.00 in steps of 0.25) in the three depth dependent matrices that could best discriminate between deleterious (destabilizing) and neutral mutations.

With the optimal parameters derived from the training set, the predictions were tested on another saturation mutagenesis set of 1534 mutants of the 101 residue long *E. coli* protein Controller of cell division or death B (CcdB) (Adkar et al., 2012; Tripathi et al., 2016) after removal of key catalytic residues (I24, I25, N95, F98, W99, G100, and I101). For the training and testing sets the crystal structures of T4 lysozyme (PDB code: 2LZM (Weaver and Matthews, 1987)) and CcdB (PDB code: 3VUB (Loris et al., 1999)) were used for depth computations. The experimental studies for T4 Lysozyme and CcdB ranked the severity of the mutant phenotype on a scale of 2–5 and 2–9 respectively. For both proteins, we considered level 2 to represent neutral (native like) mutations and all other levels to be destabilizing.

3. Results

3.1. Matrix creation and optimization

It was decided *apriori* to have a set of three 20 × 20 depth matrices, one each for exposed (E), intermediate (I) and buried (B) environments. We first determined the optimal ranges of depth values for these three matrices. As the computation of depth reports a mean value and an associated standard deviation, we decided that the minimum depth range for any of the 3 matrices should be 1.5 Å. The lower bound of the matrix corresponding to the exposed environment was set to a depth value of 2.5 Å. Its upper bound was tested in the range of 4.0 Å to 5.5 Å in steps of 0.5 Å. The lower bound of the intermediate matrix was the upper bound of the exposed matrix. It's upper bound was tested in the range of its lower bound +1.5 Å to 8.0 Å in steps of 0.5 Å. The lower bound of the buried environment matrix was the upper bound of the intermediate matrix and had no upper bound. For each combination of the three ranges, an average root mean square distance, D_M, was computed between the matrices as

$$D_M = \left\langle \sqrt{\sum_{i,j} \frac{(X_{i,j} - Y_{i,j})^2}{210}} \right\rangle$$

where X and Y are either of the three matrices E,I or B (Fig. 4) and $X_{i,j}$ is the score for substituting amino acid i for j in matrix X. The depth ranges with the highest D_M score (1.95) and hence considered optimal were (2.5–5.0 Å; 5.0–8.0 Å; > 8.0 Å). The D_M score averages

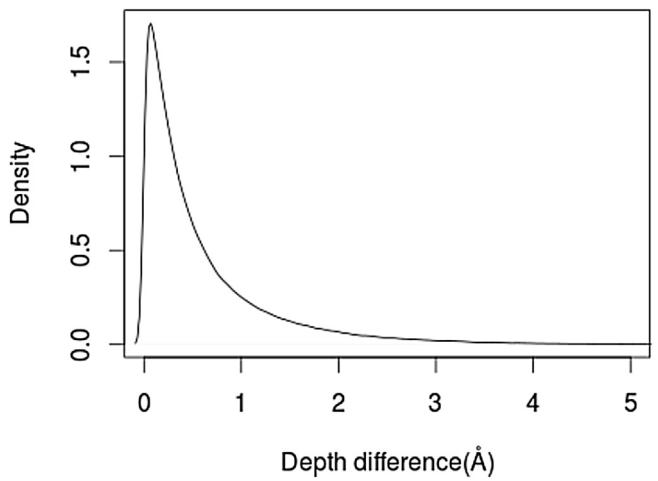

Fig. 3. Density plot of the depth difference between the aligned residues.

Please cite this article in press as: Farheen, N., et al., Depth dependent amino acid substitution matrices and their use in predicting deleterious mutations, Progress in Biophysics and Molecular Biology (2017), http://dx.doi.org/10.1016/j.pbiomolbio.2017.02.004

a

G	4.68																		
A	-0.30	2.54																	
V	-2.60	-0.22	3.78																
L	-3.01	-0.68	1.42	4.00															
I	-3.51	-0.82	3.07	2.37	4.75														
M	-1.60	-0.01	1.39	2.30	1.83	5.52													
W	-2.22	-1.75	0.11	-0.05	0.31	1.07	9.86												
F	-3.09	-1.29	0.78	1.73	1.12	1.60	3.62	6.08											
P	-1.39	-0.62	-1.07	-1.35	-1.61	-1.63	-2.70	-2.09	5.64										
S	-0.46	0.29	-0.96	-2.00	-1.83	-0.91	-2.17	-1.93	-0.62	2.99									
T	-1.69	-0.55	0.33	-0.90	0.09	-0.93	-2.01	-1.80	-1.05	1.00	3.44								
C	-2.34	0.24	0.31	0.17	-0.44	-0.80	-0.13	0.21	-2.50	-0.19	-0.90	10.71							
Y	-2.65	-1.04	-0.01	0.83	0.60	0.14	2.98	3.67	-1.89	-1.50	-1.43	0.30	6.17						
N	0.04	-0.82	-1.40	-1.44	-2.85	-0.73	-2.60	-1.47	-1.07	0.21	-0.02	-1.18	-0.60	3.56					
Q	-1.30	-0.36	-1.00	-0.96	-1.42	-0.86	-2.88	-2.16	-1.24	-0.20	0.20	-2.46	-1.28	0.07	2.92				
D	-0.63	-0.83	-2.23	-2.82	-2.99	-2.38	-3.21	-3.15	-0.69	-0.12	-0.47	-2.51	-2.65	1.01	0.00	3.61			
E	-1.71	0.20	-1.40	-1.67	-2.06	-1.83	-1.33	-2.37	-1.26	-0.58	-0.47	-2.42	-1.88	-0.42	0.77	0.88	2.76		
K	-1.43	-0.44	-1.23	-1.11	-1.29	-1.25	-1.53	-1.99	-0.85	-0.27	-0.35	-2.16	-2.07	-0.20	0.70	-0.85	0.15	2.63	
R	-1.82	-0.58	-1.09	-0.52	-1.11	-0.67	-0.86	-1.13	-1.80	-0.52	-0.68	-1.19	-1.00	-0.64	0.37	-1.45	-0.21	1.26	3.68
H	-1.15	-0.67	-0.54	-0.06	-1.26	-1.20	-1.31	0.62	-1.61	-0.22	-0.51	-0.96	1.55	0.25	0.10	-0.20	-0.70	-0.26	0.11	4.94
	G	A	V	L	I	M	W	F	P	S	T	C	Y	N	Q	D	E	K	R	H

b

G	6.43																		
A	0.58	3.52																	
V	-3.70	-0.55	2.73																
L	-4.34	-1.69	0.21	2.79															
I	-4.36	-1.46	1.75	0.73	2.98														
M	-3.57	-1.16	-0.22	1.63	0.42	4.16													
W	-3.97	-3.49	-2.62	-1.48	-2.92	-2.46	8.78												
F	-4.61	-2.32	-1.18	0.37	-0.55	-0.06	1.46	4.87											
P	-2.40	-1.15	-2.20	-2.49	-1.66	-2.39	-2.11	-2.90	8.80										
S	0.30	1.08	-2.41	-3.44	-3.19	-2.13	-3.21	-2.87	-0.75	5.00									
T	-1.64	-0.32	-0.21	-1.65	-1.39	-1.02	-2.17	-2.89	-1.81	1.84	4.84								
C	-1.84	0.31	-0.55	-1.84	-1.85	-1.71	-3.69	-1.00	-2.58	-0.27	-1.12	7.83							
Y	-3.29	-1.79	-1.43	-1.41	-1.86	-0.07	1.29	2.04	-1.71	-1.51	-1.05	-1.66	5.69						
N	-1.00	-1.84	-2.94	-3.26	-2.96	-0.76	-2.98	-3.00	-1.83	0.13	0.61	-1.15	-0.48	7.33					
Q	-1.09	-1.65	-1.36	-1.14	-1.75	-0.82	-1.80	-1.60	-1.89	0.31	0.00	-2.97	-1.69	1.54	6.44				
D	-0.96	-1.63	-4.03	-3.60	-4.43	-3.58	-4.22	-5.02	-2.68	-0.19	-1.05	-4.07	-2.34	2.50	0.23	8.00			
E	-1.96	-1.43	-2.13	-3.62	-2.73	-2.30	-2.26	-4.08	-2.75	0.44	-0.91	-3.81	-2.14	0.07	2.94	3.29	7.33		
K	-2.47	-1.49	-2.30	-2.29	-2.03	-0.30	-0.19	-2.44	-0.53	-0.60	-0.41	-2.18	-2.05	0.34	1.58	-0.85	1.92	6.95	
R	-2.05	-2.06	-2.52	-1.82	-2.60	-2.26	-0.91	-2.08	-2.23	-0.77	-0.65	-1.72	-1.67	0.34	1.80	-0.91	-0.08	3.44	7.49
H	-2.38	-2.21	-3.43	-2.80	-1.80	0.09	-1.03	-1.46	-3.57	0.26	-0.85	-2.81	0.76	1.21	1.31	-1.21	-0.99	-0.25	0.10	8.12
	G	A	V	L	I	M	W	F	P	S	T	C	Y	N	Q	D	E	K	R	H

c

G	7.08																		
A	1.39	3.81																	
V	-3.00	-0.46	2.46																
L	-4.62	-2.21	-0.70	2.34															
I	-4.52	-1.91	1.27	0.25	2.67														
M	-3.69	-0.22	-0.62	0.51	-0.52	2.96													
W	-6.62	-4.63	-4.05	-1.78	-2.29	2.53	8.77												
F	-5.39	-1.95	-1.63	0.60	-1.75	1.36	0.16	3.87											
P	-1.59	0.68	-2.12	-3.88	-3.37	-2.05	-3.75	-3.42	9.98										
S	1.36	2.09	-2.24	-2.66	-3.19	-0.98	-4.04	-3.62	0.65	5.89									
T	-1.47	0.56	-0.66	-2.09	-2.23	-0.78	-8.92	-3.01	-0.24	2.61	6.13								
C	-1.36	0.85	-1.19	-1.50	-2.48	-0.82	-5.05	-2.57	-0.49	0.75	0.98	7.94							
Y	-3.49	-1.67	-2.27	-1.46	-2.17	-0.91	0.12	2.18	-1.12	0.02	-1.84	0.07	7.11						
N	-1.73	-1.74	-2.91	-3.01	-4.19	-2.34	-4.85	-1.27	-1.85	1.31	0.88	-1.88	-2.00	8.69					
Q	-0.56	-0.81	-2.07	-1.27	-3.70	1.16	-1.45	-2.31	1.31	0.42	0.09	-3.85	0.19	-0.19	9.16				
D	-2.82	-3.77	-6.77	-6.62	-8.03	-3.10	-8.85	-5.90	-3.09	-2.02	-1.02	-3.27	0.02	1.24	0.54	11.16			
E	-3.12	-1.30	-4.53	-5.70	-5.14	-2.52	-8.00	-4.22	-1.54	-1.30	-0.78	-3.02	-4.10	-0.20	5.56	4.77	11.46		
K	-3.97	-0.45	-4.42	-4.63	-5.07	-2.76	-5.74	-5.06	-3.71	-0.89	-2.36	-4.41	-0.82	-1.71	3.12	2.54	1.92	13.07	
R	-2.39	-1.20	-4.85	-3.63	-3.87	-3.53	-2.50	-4.21	-3.45	0.65	-2.22	3.08	-1.18	-2.54	3.59	-3.99	0.62	7.42	11.50
H	-2.51	-1.89	-3.78	-1.63	-3.97	-2.57	-2.52	0.70	1.28	-1.21	-1.61	-3.59	1.27	3.69	4.18	-2.98	2.12	-2.54	1.15	9.13
	G	A	V	L	I	M	W	F	P	S	T	C	Y	N	Q	D	E	K	R	H

Fig. 4. The three symmetric 20 × 20 depth dependent substitution at depth ranges of<5 Å (a),5-8Å (b) and >8 Å (c).

over the root mean square distances of 1.44, 1.82 and 2.60 between the matrix pairs of exposed-intermediate, buried-intermediate and exposed-buried respectively.

The three depth dependent substitution matrices are all distinctly different from one another as can be gauged from their pairwise D_M values. A closer look into the pairwise matrix comparison reveals that most of the substitution scores (in the integral version) are different (Fig. 5a–c and Table 2). In all, 200 of the 210 substitution values change in the different matrices (see section 3.3 for a more description of substitution trends).

A test of accuracy of the matrix values was to create a regular substitution matrix. This composite matrix was created as

Please cite this article in press as: Farheen, N., et al., Depth dependent amino acid substitution matrices and their use in predicting deleterious mutations, Progress in Biophysics and Molecular Biology (2017), http://dx.doi.org/10.1016/j.pbiomolbio.2017.02.004

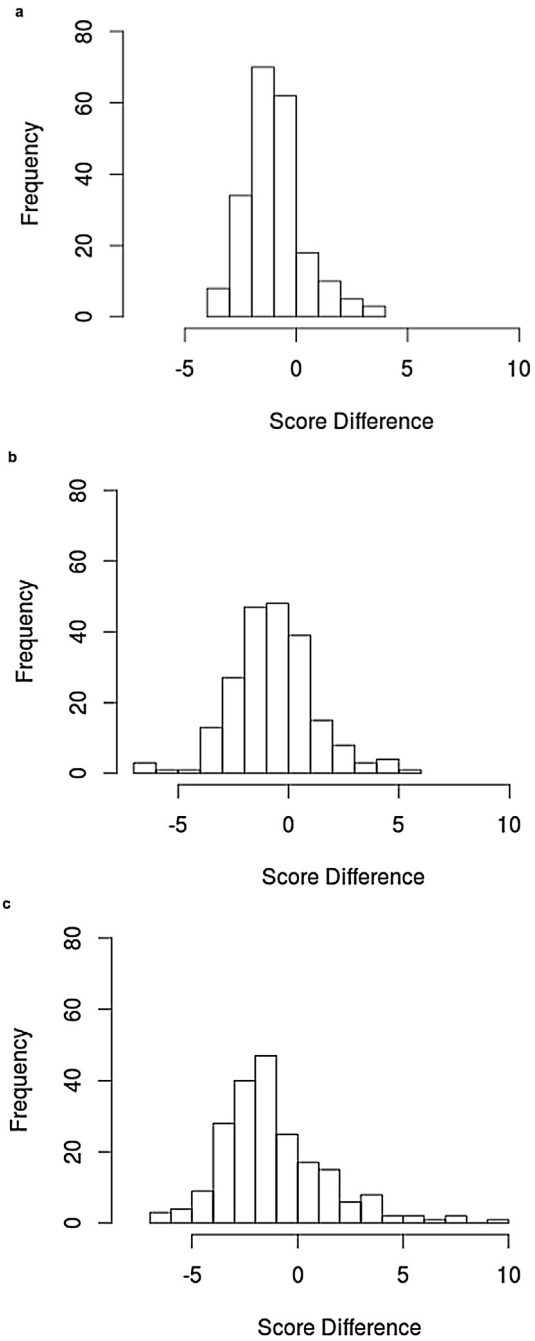

Fig. 5. Histogram of score difference of matrices between (a) Intermediate and Exposed residues (b) Buried and Intermediate residues (c) Buried and Exposed residues.

described in the methods section, only this time without taking into consideration depth levels. When our composite matrix was compared to the BLOSUM62 matrix, it was identical for 49%of the substitution values and varied by ± 1 unit in 47% of the substitution values. Here again, we believe that the differences of ± 1 are mainly caused by rounding off the matrix values. This validates that the three depth dependent matrices are stratified versions of the regular substitution matrices.

3.2. Depth conservation in alignments

The substitution matrices were created from pair wise alignments of protein structures. Over 90% of the aligned residues had depth differences of <1.5 Å, with 81% having differences of <1 Å (Fig. 3). The depth differences were examined for different types of substitutions *i.e.* polar to polar, polar to non-polar (or vice versa), and non-polar to non-polar (Table 1). Polar residues substituted by other polar residue showed the least variation in depth as these residues are predominantly found in the outer layer of the protein. Non-polar to non-polar residue substitutions showed larger depth changes in comparison. A possible reason could be that non-polar residues are present in larger proportions and interchange with one another frequently at deeper depths. In this deep environment, changes in amino acid size precipitously change their depth values. The values of depth difference for non-polar to polar substitutions (or vice versa) lie in between the two values discussed above.

3.3. Substitution trends

As discussed earlier, the relative abundances of the 20 amino acids at different depth levels are different from one another (Fig. 2). It is reasonable to expect that their substitution rates would also vary accordingly. The depth dependent substitution matrices capture these variations (Fig. 4). The substitution trends across depth levels show that the polar amino acids are easier to substitute at deeper depths while the hydrophobic amino acids show the opposite trend and get harder to substitute. It should be noted however that this higher/lower propensity of substitution is relative and at any depth amino acid self-substitutions score the highest.

Some interesting information one can extract from the depth substitution matrices are the substitution trends across depth ranges. In the matrices derived in this study there are six different types of substitution behaviors as we traverse from the outside of the protein (low depth) to the interior (high depth). Scores increase for 23 substitutions, decrease in 52 cases and remain the same in 10 substitutions. In addition to this there are 90 substitutions that have the same score for 2 consecutive depth ranges and their first/last value increase (25)/decrease (65). Some substitutions showed a trend where their values in the middle depth range had a lower or higher score as compared to scores in the other 2 ranges (35 out of 210, denoted as ∨ or ∧ in Fig. 6).

A closer look at the substitutions show that by and large the score for substituting one polar (S,T,C,Y,N,Q,D,E,K,R,H) amino acid either increases or remains the same from exposed to buried environments. This is possibly because in deeper environments substituting one polar group by another would maintain charge-charge interactions and leave no unpaired charges buried. Cysteine mutations buck this trend and are generally less favorable to mutate in deeper environments. Threonine and Serine are also less likely to be substituted by any of the larger polar (charged or uncharged) groups in deep environments. The trends for hydrophobic (G, A, V, L, I, M, W, F, P) to hydrophobic substitutions is in some sense the opposite of what is seen in polar residues. The deeper one goes into the protein the less likely it gets to substitute a non-polar group by another. This is probably because difference between the individual hydrophobic groups could contribute to substantial differences in hydrophobic packing. The trends for substituting non-polar groups by polar groups (or vice versa) get more unlikely in deeper environments. Again, there are exceptions to this - Serine, Threonine and Cysteine are more amenable to being substituted by small amino acids such as Alanine and Glycine with increasing depth. An unusual exception is the increased likelihood of substituting Tryptophan by Glutamine that gets less unfavorable

Please cite this article in press as: Farheen, N., et al., Depth dependent amino acid substitution matrices and their use in predicting deleterious mutations, Progress in Biophysics and Molecular Biology (2017), http://dx.doi.org/10.1016/j.pbiomolbio.2017.02.004

Table 1
Residue substitutions from polar to non-polar, polar to polar and non-polar to non-polar and the proportions that have depth difference of greater than 1 Å and 1.5 Å.

Type of substitution	Substitutions with depth difference >1 Å (%)	Substitutions with depth difference >1.5 Å (%)
Non-polar to non-polar	24	13
Polar to non-polar	19	10
Polar to polar	11	5

Table 2
Frequency of residues having a score difference of 0, ±1, ±2 or > ±2 between the matrices for exposed and intermediate residues, intermediate and buried residues and exposed and buried residues.

Score difference	0	±1	±2	> ±2
Exposed-Intermediate	62	88	44	16
Intermediate-Buried	48	86	42	34
Exposed-Buried	25	64	55	66

in the hydrophobic core.

The substitutions of E-H, M-Q, H-P, R-S, A-T, P-S, P-Q, C-R, C-T, A-P AND D-K that are disfavored (negative score) in the protein exterior become favored (positive score) in the hydrophobic core. The substitutions F-H, F-I, I-Y, F-V, M-V, I-M and L-Y while favored in the exterior become disfavored in the interior. In addition to these extreme cases, there are several cases of neutral mutations becoming less or more favored in different environments and vice versa. Several of the anomalous substitution behaviors could be explained away by sparse observations. Cysteine, Methionine and Tryptophan, for instance, have low abundances and hence the computation of the observed by expected substitution likelihood ratios could sometimes be erroneous. The matrix as a whole, we believe, is largely reflective of the real substitution rates between amino acids residues.

An important trend that we have not explicitly considered here is the difference in substitution rates between Cysteine (free thiol) and Cystine (disulphide bridged). In the ~3700 pairwise structure-structure alignments we have used for constructing the matrices we have very few Cystine substitutions (on average 25 substitutions of Cystine to other amino acids) and in some cases substitutions are not observed at all. For this reason, the current matrices have not differentiated between Cysteine and Cystine.

3.4. Application of the matrix to detect deleterious single point mutations

We used saturation mutagenesis data from two proteins, T4 Lysozyme and CcdB, to train and test a prediction schema for identifying destabilizing point mutations. The experimental data for both proteins described the mutagenesis data in terms of intensity of phenotype of the mutations. In the case of T4 Lysozyme the mutational sensitivity was scored on a scale of 2–5 while the range was 2–9 in the case of CcdB. For the computations described below we have taken sensitivity score of 2 to imply neutral (or native like) mutations and all other scores to imply destabilizing mutations for both proteins. The datasets hence consist of 1362 (69%) and 1258 (82%) neutral mutations and 604 (31%) and 276

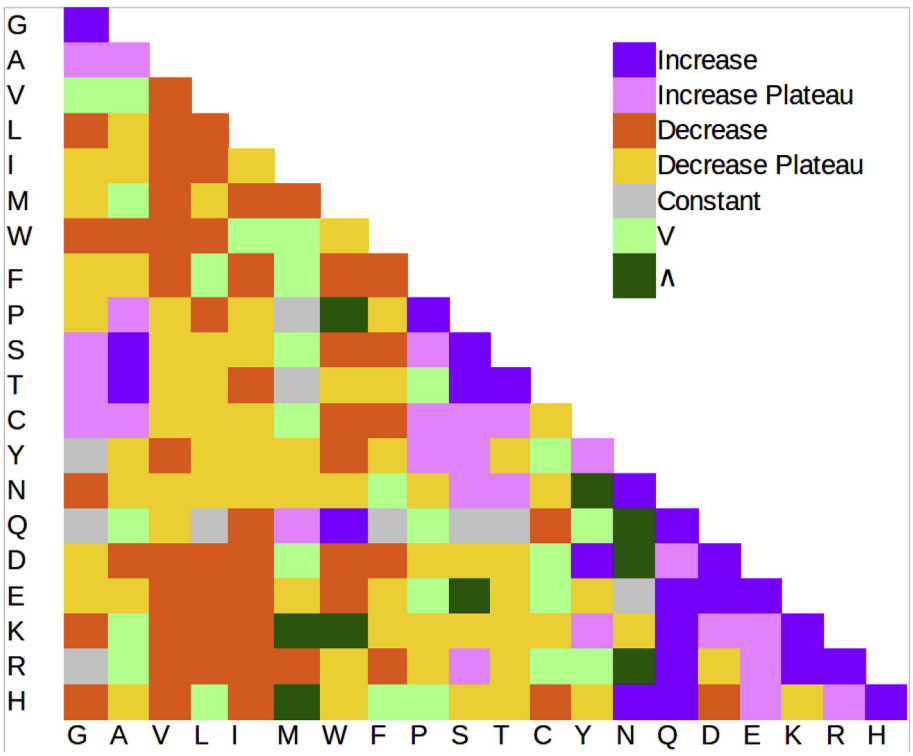

Fig. 6. Trends in substitution scores as noticed from the three depth dependent matrices. Substitutions coloured purple increase in value as depth increase. Those that decrease in value with increasing depth are coloured orange. Substitutions that have a constant value across all the depth environments are coloured grey. Substitutions that increase and then plateau and decrease and then plateau are coloured pink and yellow respectively. Substitutions that are coloured light and dark green are those whose values decrease and then increase or increase and then decrease respectively.

Please cite this article in press as: Farheen, N., et al., Depth dependent amino acid substitution matrices and their use in predicting deleterious mutations, Progress in Biophysics and Molecular Biology (2017), http://dx.doi.org/10.1016/j.pbiomolbio.2017.02.004

(18%) of destabilizing mutations in T4 Lysozyme and CcdB respectively. Our simplistic method consisted of finding threshold values in the depth substitution matrices using the training set data that would best distinguish between neutral and destabilizing mutations. The thresholds were found by a grid search that varied the threshold value in the range −3 to −0.25 in steps of 0.25 over all three matrices. The optimal threshold values (−3, −0.25 and −0.25 for the <5 Å, 5–8 Å and >8 Å depth matrices respectively) were then applied to evaluate the efficacy of the method over the testing set data. The accuracy of our binary classification method, and those of other methods compared to ours, was measured in terms of Sensitivity, Specificity, Precision, Accuracy, f1 and MCC(Matthews, 1975) (Weaver and Matthews, 1987).

Our depth dependent substitution matrix (FADHM) method was compared to other popular methods including Automute (Masso and Vaisman, 2010), DUET (Pires et al., 2014b), I-mutant (Capriotti et al., 2005), SuSPect (Yates et al., 2014), msCSM (Topham et al., 1997) and SDM(Worth et al., 2011) which predict if mutations are destabilizing (Table 3). Our precision values (60% and 44% for T4 lysozyme and CcdB respectively) and specificity values (85% and 78% for T4 lysozyme and CcdB respectively) were either the highest or comparable to that of the other methods. Consistently, in both the training and testing sets FADHM has the best accuracy (75% and 78%), f1 (0.55 and 0.57) and MCC values (0.38 and 0.48). The next best methods for theT4 lysozyme and CcdB datasets have MCC values of 0.30 (I-mutant) and 0.40 (DUET) respectively. Predictions of destabilizing mutations by our simple method outperform other sophisticated algorithms.

To check the robustness of our results we repeated the accuracy computations 10 times for both the training and testing sets, this time considering only a randomly selected 40% subset of the data. These tests showed that the average MCC value for T4 Lysozyme and CcdB were 0.39 with a standard deviation of 0.03 and 0.48 with a standard deviation of 0.02 respectively.

4. Discussion

In earlier studies we had established the utility of the residue depth measure to concisely describe local environment. The depth measure has been successfully used for diverse applications including, but not exhaustive, finding small molecule binding sites on proteins, predicting what single point mutations would yield temperature sensitive mutations and estimating the pK_as of ionizable amino acids. In this study we have used the depth measure in conjunction with the knowledge that the relative abundances of different amino acids change with protein environments. This suggests that the substitution rates of amino acids would also be different at different depths. The 3 depth dependent substitution matrices were hence created.

We arbitrarily chose to create a set of three matrices to represent the substitutions in expose, intermediate and buried environments. The depth values (5 Å and 8 Å) that demarcated the boundaries between these 3 classes were obtained by attempting to maximize the differences between the matrices. The resulting matrices are quite different from one another and show the difference in the substitutions likelihoods in different environments. We observed 6 different substitution trends in pairwise residue substitutions across different environments. The patterns include substitutions whose values remained unchanged, increased or decreased monotonically, increased/decreased and then plateaued, increase and then decreased or vice versa. Only 10 of the 210 substitutions remained unchanged across all three environments. The matrices show many expected trends such as how replacing a hydrophobic residue with a polar one in the buried environments is generally unfavourable. There were many surprising substitutions trends where the intermediate region was the most favoured in comparison to buried and exposed environments. Some of these trends could be artifacts of low abundance of residues such as Methionine, Cysteine and Tryptophan. The other such trends indicate that the matrices were able to capture some of the nuances of residue preferences and substitutions across different environments.

We tested the matrices for their ability to detect mutations that lead to protein instability. Saturation mutagenesis data from T4 Lysozyme and CcdB were used as the training and testing sets respectively. Mutations/substitutions were considered as destabilizing if the substitution score (native to mutant) was less than −3.00, −0.25 and −0.25 in exposed, intermediate and buried environments respectively. Our somewhat simplistic approach outperformed other popular methods, some of which use machine learning rigorously. Of the 276 deleterious mutation in the CcdB test set, we accurately identified 220 while the next best method identified only ~160. Our method, and the others, produce a large number of false positives and hence the somewhat modest overall performance (MCC of 0.38 on the training set and 0.48 on the testing set). In comparison to the others, our method has low sensitivity, is comparable in terms of specificity and precision but clearly outperforms in accuracy, f1 and MCC values.

We believe that these depth dependent substitution matrices

Table 3
Prediction performance comparison of different prediction techniques on (a) training set(T4 lysozyme) (b) testing set(CcdB). Maximum performance of each value measure is indicated in bold. * FADHM is Amino acid DeptH substitution Matrices.

Technique	Sensitivity (%)	Specificity (%)	Precision (%)	Accuracy (%)	F1	MCC
a)						
I-mutant	56	75	50	69	0.53	0.30
SuSPect	**75**	53	41	60	0.53	0.26
Automute	70	54	41	59	0.52	0.23
mCSM	60	70	26	68	0.37	0.22
SDM	58	73	28	71	0.38	0.24
DUET	61	70	27	69	0.38	0.24
FADHM*	51	**85**	**60**	**75**	**0.55**	**0.38**
b)						
I-mutant	64	78	44	73	0.52	0.36
SuSPect	**99**	21	22	35	0.36	0.21
Automute	76	49	30	55	0.43	0.21
mCSM	68	76	44	74	0.54	0.39
SDM	54	**81**	45	75	0.49	0.33
DUET	67	78	**46**	76	0.54	0.40
FADHM	80	78	44	**78**	**0.57**	**0.48**

Please cite this article in press as: Farheen, N., et al., Depth dependent amino acid substitution matrices and their use in predicting deleterious mutations, Progress in Biophysics and Molecular Biology (2017), http://dx.doi.org/10.1016/j.pbiomolbio.2017.02.004

are important in describing the internal environments of proteins. Further developments of these potentials could include the creation of asymmetric substitution matrices as the relative abundances of different amino acids in the different environments vary. These matrices should be able to improve the accuracy in aligning distantly related homologues with one another. This is the first of what we expect to be a series of studies to learn from substitution likelihoods in different protein environments.

Funding

This work was supported by a Wellcome trust-DBT India alliance senior fellowship.

Conflict of interest

None declared.

Acknowledgements

M.S.M would like to acknowledge the Wellcome trust-DBT India alliance for a senior fellowship, NS holds a CSIR SPM fellowship and NF would like to acknowledge an INSPIRE-SHE fellowship. The authors would like to thank Neelesh Soni for discussions and critical comments.

References

Abascal, F., Posada, D., Zardoya, R., 2007. MtArt: a new model of amino acid replacement for Arthropoda. Mol. Biol. Evol. 24, 1–5. http://dx.doi.org/10.1093/molbev/msl136.

Adachi, J., Hasegawa, M., 1996. Model of amino acid substitution in proteins encoded by mitochondrial DNA. J. Mol. Evol. 42, 459–468 doi:8642615.

Adkar, B.V., Tripathi, A., Sahoo, A., Bajaj, K., Goswami, D., Chakrabarti, P., Swarnkar, M.K., Gokhale, R.S., Varadarajan, R., 2012. Protein model discrimination using mutational sensitivity derived from deep sequencing. Structure 20, 371–381. http://dx.doi.org/10.1016/j.str.2011.11.021.

Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S., Sunyaev, S.R., 2010. A method and server for predicting damaging missense mutations. Nat. Methods 7, 248–249. http://dx.doi.org/10.1038/nmeth0410-248.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. J. Mol. Biol. 215, 403–410. http://dx.doi.org/10.1016/S0022-2836(05)80360-2.

Anfinsen, C.B., 1973. Principles that govern the folding of protein chains. Sci. (80-.) 181, 223–230. http://dx.doi.org/10.1126/science.181.4096.223.

Arvestad, L., 2006. Efficient methods for estimating amino acid replacement rates. J. Mol. Evol. 62, 663–673. http://dx.doi.org/10.1007/s00239-004-0113-9.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E., 2000. The protein data bank. Nucleic Acids Res. 28, 235–242. http://dx.doi.org/10.1093/nar/28.1.235.

Braberg, H., Webb, B.M., Tjioe, E., Pieper, U., Sali, A., Madhusudhan, M.S., 2012. Salign: a web server for alignment of multiple protein sequences and structures. Bioinformatics 28, 2072–2073. http://dx.doi.org/10.1093/bioinformatics/bts302.

Capriotti, E., Fariselli, P., Casadio, R., 2005. I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. Nucleic Acids Res. 33 http://dx.doi.org/10.1093/nar/gki375.

Chakravarty, S., Varadarajan, R., 1999. Residue depth: a novel parameter for the analysis of protein structure and stability. Structure 7, 723–732. http://dx.doi.org/10.1016/S0969-2126(99)80097-5.

Chenna, R., Sugawara, H., Koike, T., Lopez, R., Gibson, T.J., Higgins, D.G., Thompson, J.D., 2003. Multiple sequence alignment with the Clustal series of programs. Nucleic Acids Res. 31 (13), 3497–3500. http://dx.doi.org/10.1093/nar/gkg500.

Chothia, C., 1992. Proteins. One thousand families for the molecular biologist. Nature 357, 543–544. http://dx.doi.org/10.1038/357543a0.

Dayhoff, M., Schwartz, R., 1978. A model of evolutionary change in proteins. Atlas protein Seq. Struct. 345–352 doi:10.1.1.145.4315.

DeLano, W.L., 2002. The PyMOL Molecular Graphics System. wwwpymolorg Version 1.http://www.pymol.org. Schrödinger LLC. citeulike-article-id:240061.

Dimmic, M.W., Rest, J.S., Mindell, D.P., Goldstein, R.A., 2002. rtREV: an amino acid substitution matrix for inference of retrovirus and reverse transcriptase phylogeny. J. Mol. Evol. 55, 65–73. http://dx.doi.org/10.1007/s00239-001-2304-y.

Fiser, a, Fiser, A., Do, R.K., Do, R.K., Sali, A., Sali, A., 2000. Modeling of loops in protein structures. Protein Sci. 9, 1753–1773. http://dx.doi.org/10.1110/ps.9.9.1753.

Goldman, N., Thorne, J.L., Jones, D.T., 1998. Assessing the impact of secondary structure and solvent accessibility on protein evolution. Genetics 149, 445–458. http://dx.doi.org/10.1093/molbev/msl086.

Henikoff, S., Henikoff, J.G., 1992. Amino acid substitution matrices from protein blocks. Proc. Natl. Acad. Sci. U. S. A. 89, 10915–10919. http://dx.doi.org/10.1073/pnas.89.22.10915.

Johnson, M.S., Overington, J.P., Blundell, T.L., 1993. Alignment and searching for common protein folds using a data bank of structural templates. J. Mol. Biol. 231, 735–752. http://dx.doi.org/10.1006/jmbi.1993.1323.

Jones, D.T., Taylor, W.R., Thornton, J.M., 1994. A mutation data matrix for transmembrane proteins. FEBS Lett. 339, 269–275. http://dx.doi.org/10.1016/0014-5793(94)80429-X.

Koshi, J.M., Goldstein, R.A., 1995. Context-dependent optimal substitution matrices. Protein Eng. Des. Sel. 8, 641–645. http://dx.doi.org/10.1093/protein/8.7.641.

Lartillot, N., Philippe, H., 2004. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Mol. Biol. Evol. 21, 1095–1109. http://dx.doi.org/10.1093/molbev/msh112.

Lee, B., Richards, F.M., 1971. The interpretation of protein structures: estimation of static accessibility. J. Mol. Biol. 55 http://dx.doi.org/10.1016/0022-2836(71)90324-X.

Liu, S., Zhang, C., Liang, S., Zhou, Y., 2007. Fold recognition by concurrent use of solvent accessibility and residue depth. Proteins Struct. Funct. Genet. 68, 636–645. http://dx.doi.org/10.1002/prot.21459.

Loris, R., Dao-Thi, M.H., Bahassi, E.M., Van Melderen, L., Poortmans, F., Liddington, R., Couturier, M., Wyns, L., 1999. Crystal structure of CcdB, a topoisomerase poison from E. coli. J. Mol. Biol. 285, 1667–1677. http://dx.doi.org/10.1006/jmbi.1998.2395.

Lüthy, R., McLachlan, a D., Eisenberg, D., 1991. Secondary structure-based profiles: use of structure-conserving scoring tables in searching protein sequence databases for structural similarities. Proteins 10, 229–239. http://dx.doi.org/10.1002/prot.340100307.

Madhusudhan, M.S., Webb, B.M., Marti-Renom, M.A., Eswar, N., Sali, A., 2009. Alignment of multiple protein structures based on sequence and structure features. Protein Eng. Des. Sel 22, 569–574. http://dx.doi.org/10.1093/protein/gzp040.

Masso, M., Vaisman, I.I., 2010. AUTO-MUTE: web-based tools for predicting stability changes in proteins due to single amino acid replacements. Protein Eng. Des. Sel. 23, 683–687. http://dx.doi.org/10.1093/protein/gzq042.

Matthews, B.W., 1975. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. BBA - Protein Struct. 405, 442–451. http://dx.doi.org/10.1016/0005-2795(75)90109-9.

Mehta, P.K., Heringa, J., Argos, P., 1995. A simple and fast approach to prediction of protein secondary structure from multiply aligned sequences with accuracy above 70%. Protein Sci. 4, 2517–2525. http://dx.doi.org/10.1002/pro.5560041208.

Ng, P.C., Henikoff, S., 2003. SIFT: predicting amino acid changes that affect protein function. Nucleic Acids Res. 31, 3812–3814. http://dx.doi.org/10.1093/nar/gkg509.

Overington, J., Johnson, M.S., Sali, a, Blundell, T.L., 1990. Tertiary structural constraints on protein evolutionary diversity: templates, key residues and structure prediction. Proc. Biol. Sci. http://dx.doi.org/10.1098/rspb.1990.0077.

Pintar, A., Carugo, O., Pongor, S., 2003a. Atom depth as a descriptor of the protein interior. Biophys. J. 84, 2553–2561. http://dx.doi.org/10.1016/S0006-3495(03)75060-7.

Pintar, A., Carugo, O., Pongor, S., 2003b. Atom depth in protein structure and function. Trends biochem. Sci. http://dx.doi.org/10.1016/j.tibs.2003.09.004.

Pires, D.E.V., Ascher, D.B., Blundell, T.L., 2014a. MCSM: predicting the effects of mutations in proteins using graph-based signatures. Bioinformatics 30, 335–342. http://dx.doi.org/10.1093/bioinformatics/btt691.

Pires, D.E.V., Ascher, D.B., Blundell, T.L., 2014b. DUET: a server for predicting effects of mutations on protein stability using an integrated computational approach. Nucleic Acids Res. 42 http://dx.doi.org/10.1093/nar/gku411.

Rennell, D., Bouvier, S.E., Hardy, L.W., Poteete, A.R., 1991. Systematic mutation of bacteriophage T4 lysozyme. J. Mol. Biol. 222 http://dx.doi.org/10.1016/0022-2836(91)90738-R.

Rice, D.W., Eisenberg, D., 1997. A 3D-1D substitution matrix for protein fold recognition that includes predicted secondary structure of the sequence. J. Mol. Biol. 267, 1026–1038. http://dx.doi.org/10.1006/jmbi.1997.0924.

Sali, A., Blundell, T.L., 1993. Comparative protein modelling by satisfaction of spatial restraints. J. Mol. Biol. 234, 779–815. http://dx.doi.org/10.1006/jmbi.1993.1626.

Shi, J., Blundell, T.L., Mizuguchi, K., 2001. FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. J. Mol. Biol. 310, 243–257. http://dx.doi.org/10.1006/jmbi.2001.4762.

Tan, K.P., Varadarajan, R., Madhusudhan, M.S., 2011. DEPTH: a web server to compute depth and predict small-molecule binding cavities in proteins. Nucleic Acids Res. 39 http://dx.doi.org/10.1093/nar/gkr356.

Tan, K.P., Nguyen, T.B., Patel, S., Varadarajan, R., Madhusudhan, M.S., 2013. Depth: a web server to compute depth, cavity sizes, detect potential small-molecule ligand-binding cavities and predict the pKa of ionizable residues in proteins. Nucleic Acids Res. 41 http://dx.doi.org/10.1093/nar/gkt503.

Tan, K.P., Khare, S., Varadarajan, R., Madhusudhan, M.S., 2014. TSpred: a web server for the rational design of temperature-sensitive mutants. Nucleic Acids Res. 42 http://dx.doi.org/10.1093/nar/gku319.

Thorne, J.L., Goldman, N., Jones, D.T., 1996. Combining protein evolution and

secondary structure. Mol. Biol. Evol. 13, 666–673. http://dx.doi.org/10.1093/oxfordjournals.molbev.a025627.

Topham, C.M., Srinivasan, N., Blundell, T.L., 1997. Prediction of the stability of protein mutants based on structural environment-dependent amino acid substitution and propensity tables. Protein Eng. 10, 7–21. http://dx.doi.org/10.1093/protein/10.1.7.

Tripathi, A., Gupta, K., Khare, S., Jain, P.C., Patel, S., Kumar, P., Pulianmackal, A.J., Aghera, N. and Vardarajan R., 2016. Molecular determinants of mutant phenotypes, inferred from saturation mutagenesis data. Mol. Biol. Evol. 1–35. http://dx.doi.org/10.1093/molbev/msw182.

Wako, H., Blundell, T.L., 1994. Use of AA env-dependent substitution tables and conf propensities in struc prediction from aligned sequences of homologous proteins. II. Second. Struc. J. Mol. Biol. 238, 693–708. http://dx.doi.org/10.1006/jmbi.1994.1330.

Wang, G., Dunbrack, R.L., 2003. PISCES: a protein sequence culling server. Bioinformatics 19, 1589–1591. http://dx.doi.org/10.1093/bioinformatics/btg224.

Wang, Z., Moult, J., 2001. SNPs, protein structure, and disease. Hum. Mutat. 17, 263–270. http://dx.doi.org/10.1002/humu.22.

Weaver, L.H., Matthews, B.W., 1987. Structure of bacteriophage T4 lysozyme refined at 1.7 Å resolution. J. Mol. Biol. 193, 189–199. http://dx.doi.org/10.1016/0022-2836(87)90636-X.

Wolynes, P.G., Onuchic, J.N., Thirumalai, D., 1995. Navigating the folding routes. Science 267, 1619–1620. http://dx.doi.org/10.1126/science.7886447.

Worth, C.L., Preissner, R., Blundell, T.L., 2011. SDM–a server for predicting effects of mutations on protein stability and malfunction. Nucleic Acids Res. 39, W215–W222. http://dx.doi.org/10.1093/nar/gkr363.

Yates, C.M., Filippis, I., Kelley, L.A., Sternberg, M.J.E., 2014. SuSPect: enhanced prediction of single amino acid variant (SAV) phenotype using network features. J. Mol. Biol. 426, 2692–2701. http://dx.doi.org/10.1016/j.jmb.2014.04.026.

Zhou, H., Zhou, Y., 2005. Fold recognition by combining sequence profiles derived from evolution and from depth-dependent structural alignment of fragments. Proteins Struct. Funct. Genet. 58, 321–328. http://dx.doi.org/10.1002/prot.20308.

Please cite this article in press as: Farheen, N., et al., Depth dependent amino acid substitution matrices and their use in predicting deleterious mutations, Progress in Biophysics and Molecular Biology (2017), http://dx.doi.org/10.1016/j.pbiomolbio.2017.02.004