BIOINFORMATICS

Vol. 00 no. 00 2008
Pages 1–3

A Hierarchy of Delaunay Tessellation-based Scoring Functions: Supplement I – Solubility Mutation Dataset

Andrew H. Fowler†, Bala Krishnamoorthy*, Kelly Stratton*, and Christopher Deutsch‡

†- Schweitzer Engineering Laboratories, Pullman; * Department of Mathematics, Washington State University; ‡ Department of Chemistry, Portland State University.

We present a dataset of 67 mutants along with information on the changes to the solubility of the proteins resulting from the mutations in Tables (1) and (2). These mutants were assembled from several articles. We have 20 mutants from the work of Trevino et al. (2007), 26 of the mutants from the dataset considered by Idicula-Thomas and Balaji (2005), 21 mutants assembled from the paper of Sim and Sim (1999) and references therein, and six mutants considered by Maxwell et al. (1999). We did not include all the mutants reported in these papers, because of the unavailability of crystal structures of the WT proteins in certain cases.

We also report the results of predicting the effects of mutagenesis on solubility of the proteins in this dataset using the three-body scoring function based on the Delaunay tessellation (DT) of the WT structure (details of the scoring function are presented in the main paper). The accuracy of prediction of the DT-based scoring function on the first subset (1-20) of mutants is 80% (16 out of 20), on the second subset (21-46) of mutants is 57.7 % (15 out of 26), on the third subset (47-61) of mutants is 73.3% (11 out of 15), and that on the fourth subset (62-67) of mutants is 83.3% (5 out of 6). The overall accuracy is 70.2% (47 out of 67). While this level of accuracy is comparable to some of the highest accuracies reported for similar methods (Smialowski et al., 2007), it is more significant to note that our scoring function is developed independent of learning from any of the solubility mutants available. We must mention, though, that most current methods for predicting solubility mutagenesis use only the sequence information, and not much information about the 3D structure of the WT (or mutant, if available). Since our scoring function is mainly based on structure, the higher accuracy of prediction may not be surprising.

We are currently searching the literature for other similar mutants (with data on associated solubility changes) that could be added to the dataset of solubility mutations. Once we have a large-enough dataset, we plan to develop more accurate DT-based scoring functions for solubility mutagenesis by training from a subset of the mutants. At the same time, it is interesting to note that the current dataset is already a large collection of solubility mutants as compared to previously reported sets. Idicula-Thomas and Balaji (2005) presented a set of 34 mutations (most of which are absorbed into our dataset), while Smialowski et al. (2007) presented a comparison of the performances of various scoring functions for predicting solubility mutagenesis on a dataset of 64 mutants.

REFERENCES

Idicula-Thomas, S. and Balaji, P. V. (2005) Understanding the relationship between the primary structure of proteins and its propensity to be soluble on overexpression in Escherichia coli. *Protein Sci*, **14**(3), 582–592.

Maxwell, K. L., Mittermaier, A. K., Forman-Kay, J. D., and Davidson, A. R. (1999) A simple in vivo assay for increased protein solubility. *Protein Science*, **8**(9), 1908–1911.

Sim, J. and Sim, T. (1999) Amino acid substitutions affecting protein solubility: high level expression of streptomyces clavuligerus isopenicillin N synthase in escherichia coli. *Journal of Molecular Catalysis B: Enzymatic*, **6**(3), 133–143.

Smialowski, P., Martin-Galiano, A. J., Mikolajika, A., Girschick, T., Holak, T. A., and Frishman, D. (2007) Protein solubility: sequence based prediction and experimental verification. *Bioinformatics*, **23**(19), 2536–2542.

Trevino, S. R., Scholtz, J., and Pace, C. (2007) Amino acid contribution to protein solubility: Asp, Glu, and Ser contribute more favorably than the other hydrophilic amino acids in RNase Sa. *Journal of Molecular Biology*, **366**(2), 449–460.

Table 1. First 47 entries from the dataset of solubility mutants. Entries in the column "Sol" indicate whether, as a result of the mutation(s), the solubility increased (indicated by 1), decreased (-1), or remained unchanged or was similar to that of WT (0). "Score" gives the DT-based score for the mutation (see main paper for details). The column "Pred" indicates a 1 if the prediction was correct (increase, decrease, or no change in solubility), and 0 if wrong. Information on mutants 1-20 was collected from the work of Trevino et al. (2007), while mutants 21-46 were assembled from the work of Idicula-Thomas and Balaji (2005), and references therein.

#	Protein	PDB chain	Mutations	Sol	Score	Pred
1	Rnase Sa	1RGG A	THR 76 ASP	1	-1.388	0
2			THR 76 ARG	1	6.948	1
3			THR 76 GLU	1	1.076	1
4			THR 76 SER	1	0.4874	1
5			THR 76 LYS	1	8.849	1
6			THR 76 GLY	1	-1.328	0
7			THR 76 ALA	1	1.111	1
8			THR 76 HIS	1	1.611	1
9			THR 76 ASN	1	2.425	1
10			THR 76 THR	0	0	1
11			THR 76 GLN	0	6.929	0
12			THR 76 PRO	-1	3.542	0
13			THR 76 CYS	-1	-13.93	1
14			THR 76 MET	-1	-5.481	1
15			THR 76 VAL	-1	-9.654	1
16			THR 76 LEU	-1	-6.435	1
17			THR 76 ILE	-1	-12.02	1
18			THR 76 TYR	-1	-7.264	1
19			THR 76 PHE	-1	-12.03	1
20			THR 76 TRP	-1	-6.46	1
21	Human HIV type 1 integrase	1BIZ A	TRP 131 ALA	1	0.4892	1
22			VAL 165 LYS	1	11.62	1
23	aldehyde dehydrogenase (human)	1NZX A	CYS 19 TYR	-1	14.4	0
24			ALA 104 THR	1	6.233	1
25			TYR 203 HIS	1	2.94	1
26	Human galactokinase	1WUU A	PRO 1 28 THR	-1	0.8014	0
27			VAL 32 MET	-1	-3.382	1
28			GLY 36 ARG	-1	0	0
29			THR 288 MET	-1	-0.4242	1
30			ALA 384 PRO	-1	0.7147	0
31	Basic fibroblast growth factor	1FGA A	CYS 70 SER	-1	-2.989	1
32			CYS 26 SER	-1	-3.014	1
33			CYS 93 SER	-1	8.466	0
34	Colicin A	1COL A	TRP 140 PHE	1	-2.906	0
35			TRP 140 LYS	-1	-1.777	1
36			TRP 140 LEU	-1	-4.013	1
37			TRP 140 CYS	-1	-14.45	1
38			TRP 86 PHE, TRP 140 PHE	-1	-1.951	1
39			TRP 130 PHE, TRP 140 PHE	-1	-1.732	1
40			TRP 86 PHE, TRP 130 PHE, TRP 140 PHE	-1	-1.254	1
41	Galactokinase	1WUU A	PRO 28 LYS	-1	1.986	0
42			HIS 44 TYR	-1	9.454	0
43			ARG 68 CYS	-1	-10.55	1
44			GLY 346 SER	-1	4.027	0
45			GLY 349 SER	-1	1.966	0
46			ALA 198 VAL	-1	0.3307	0

Table 2. Remaining 21 entries from the dataset of solubility mutants, continued from Table 1. Information on mutants 47-61 was collected from the work of Sim and Sim (1999) and references therein, while mutants 62-67 were collected from the work of Maxwell et al. (1999). Notice the Score for mutant 66 – even though it is negative, and hence could indicate a decrease in solubility, its magnitude is lower than the chosen cut-off value of 0.1%. As such, we record this prediction as an incorrect one.

#	Protein	PDB	Mutations	Sol	Score	Pred
47	Human alpha-1 proteinase inhibitor	8API A	MET 351 GLU, MET 358 ARG	1	6.854	1
48			THR 345 LEU, MET 358 ARG	-1	5.91	0
49			MET 358 LEU	-1	-0.3838	1
50	Human Interleukin 1 Beta	9ILB A	LYS 97 ARG	1	2.525	1
51			LYS 97 GLY	-1	-5.205	1
52			LYS 97 VAL	-1	-10.05	1
53	Colicin A	1COL A	TRP 140 LYS	-1	-1.777	1
54			TRP 140 LEU	-1	-4.013	1
55			TRP 140 CYS	-1	-14.45	1
56			LYS 113 PHE, TRP 140 LYS	1	-8.139	0
57			LYS 113 PHE, TRP 140 LEU	1	-8.705	0
58			LYS 113 PHE, TRP 140 CYS	1	-14.12	0
59	Human Interleukin 1 Beta	9ILB A	LEU 10 ASN	-1	-17.37	1
60			LEU 10 ASP	-1	-23.68	1
61			LEU 10 THR	-1	-11.52	1
62	HIV integrase	1BIZ A	LYS 185 PHE	-1	-9.711	1
63			LYS 185 ILE	-1	-12.16	1
64			LYS 185 VAL	-1	-13.9	1
65			LYS 185 LEU	-1	-11.86	1
66			LYS 185 ASN	-1	-0.08249	0
67			LYS 185 ASP	-1	-1.758	1