Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Prediction Model**

Supplementary material

Sequence derived parameters	Enzyme		Non-Enzyme		Sequence derived parameters	Enzyme		Non-Enzyme	
	Max	Min	Max	Min		Max	Min	Max	Min
Molecular Weight	0.207588	0.00182	0.20947	0.00419	N_DayhoffStat	0.1671	0.0987	0.2114	0.1078
Average Residue	0.11811	0.09159	0.1209	0.09186	P_Mole %	0.9572	0.3450	3.6556	0.5680
Isoelectric Point	0.104656	0.0427	0.1288	0.03857	P_DayhoffStat	0.1841	0.0089	0.703	0.02908
Extinction Coefficient	0.29032	0.019	0.33257	0.027	Q_Mole %	0.585	0.0871	1.5106	0.1098
Extinction Coefficient (1 mg/ml)	0.275	0.024	0.376	0.036	Q_DayhoffStat	0.15	0.0098	0.3873	0.0129
Improablity / Proability inclusion bodies	0.928	0.494	0.979	0.41	R_Mole %	1.0682	0.0088	2.1256	0.0187
A_Mole %	0.18828	0.02881	0.21186	0.03	R_DayhoffStat	0.218	0.02389	0.434	0.0452
A_DayhoffStat	0.2189	0.0335	0.2464	0.045	S_Mole %	0.9035	0.1796	2.2034	0.0012
B_Mole %	0.1989	0.0017	0.0902	0.0011	S_DayhoffStat	0.1291	0.0257	0.3148	0.0389
B_DayhoffStat	0.0292	0.001	0.0109	0.0009	T_Mole %	1.0497	0.3091	1.4352	0.1203
C_Mole %	1	0.00659	2.0339	0.0089	T_DayhoffStat	0.1721	0.0507	0.2353	0.0092
C_DayhoffStat	0.3448	0.02154	0.7013	0.0154	V_Mole %	0.15	0.04484	0.17647	0.0289
D_Mole %	0.8147	0.0154	1.206	0.0015	V_DayhoffStat	0.2273	0.0679	0.2674	0.0546
D_DayhoffStat	0.1481	0.0152	0.2193	0.0652	W_Mole %	0.4598	0.00245	0.4839	0.0254
E_Mole %	1.018	0.0147	1.8615	0.0254	W_DayhoffStat	0.3537	0.0021	0.3722	0.0215
E_DayhoffStat	0.1697	0.0215	0.3102	0.0145	X_Mole %	0.4562	0.025	0.3262	0.0254
F_Mole %	0.9195	0.1277	1.0044	0.0596	X_DayhoffStat	0.5263	0.0562	0.3215	0.025
F_DayhoffStat	0.2554	0.0355	0.279	0.0101	Y_Mole %	0.6135	0.0159	2.4615	0.0521
G_Mole %	0.25	0.00769	0.36923	0.00503	Y_DayhoffStat	0.1804	0.0154	0.724	0.00987
G_DayhoffStat	0.2976	0.0092	0.4396	0.006	Z_Mole %	0.2222	0.0089	0.3262	0.0154
H_Mole %	0.6513	0.00894	1.0271	0.021	Z_DayhoffStat	0.894	0.1256	0.265	0.03652
H_DayhoffStat	0.3257	0.0456	0.5136	0.0598	Tiny Mole %	0.6	0.15569	0.6389	0.16239
I_Mole %	1	0.2077	1.0377	0.0089	Small Mole %	0.75	0.4012	0.77119	0.32479
I_DayhoffStat	0.2222	0.0462	0.2306	0.0564	Aliphatic Mole %	0.31481	0.14808	0.32903	0.02542
K_Mole %	1.018	0.0591	2.0455	0.00115	Aromatic Mole %	0.24521	0.04918	0.29231	0.08541
K_DayhoffStat	0.1542	0.00213	0.3099	0.0002	Non-polar Mole %	0.85	0.45521	0.86154	0.31818
L_Mole %	0.19444	0.03139	0.19101	0.0321	Polar Mole %	0.54479	0.15	0.68182	0.13846
L_DayhoffStat	0.2628	0.0424	0.2581	0.0021	Charged Mole %	0.33533	0.05	0.46986	0.01389
M_Mole %	0.5169	0.0456	1.2346	0.0268	Basic Mole %	0.17365	0.05	0.31624	0.00926
M_DayhoffStat	0.3041	0.0154	0.7262	0.0158	Acidic Mole %	0.16168	0.00897	0.25	0.0154
N_Mole %	0.7186	0.1200	0.9091	0.2300					

Table 1: 61 'Pepstat (EMBOSS)' primary sequence descriptors used in the study. The parameters are scaled down by appropriate scaling values.

ISSN 0973-2063
Bioinformation 2(3):107-112 (2007)

5-fold cross validation	Accuracy	Specificity	Sensitivity	MCC	Q(Pred)	Prediction range (enzymes)	Prediction range (non-enzymes)
(a) using sequence derived features (PEPSTAT)							
C1	0.8947	1.00	0.8271	0.8072	100	0.9626-1.00	0.00-0.5340
C2	0.7969	0.7671	0.8333	0.5979	74.62	0.9579-1.00	0.00-0.6758
C3	0.7142	0.6794	0.7636	0.4364	76.68	0.9257-1.00	0.00-0.8786
C4	0.7443	0.6666	0.9495	0.5490	52.28	0.9692-1.00	0.00-0.8586
C5	0.7894	0.7934	0.8545	0.5891	70.14	0.9048-1.00	0.00-0.8236
Mean	0.7879 ± 0.0686	0.7713 ± 0.1339	0.8448 ± 0.0673	0.5959 ± 0.1345	74.734 ± 17.084		
(b) using PSSM matrix (PSI BLAST)							
C1	0.8230	0.7641	0.9158	0.6628	71.15	0.9237-0.9559	0.2180-0.2205
C2	0.8717	0.8148	0.9538	0.7560	78.13	0.9357-0.9443	0.3921-0.6006
C3	0.8521	0.8072	0.9123	0.7118	77.91	0.9061-0.9156	0.1626-0.7521
C4	0.7567	0.6988	0.8624	0.5368	61.09	0.9255-0.9272	0.3239-0.5133
C5	0.7153	0.6485	0.8911	0.4821	49.05	0.9123-0.9343	0.3005-0.4183
Mean	0.8037 ± 0.0659	0.7466 ± 0.0717	0.9070 ± 0.0337	0.6299 ± 0.1164	67.466 ± 12.411		

Table 2: Results of enzymes / non-enzymes prediction methods, using five fold cross validation

Equations used in this article:

Accuracy of the prediction methods

$$Q_{ACC} = \frac{P+N}{T} \text{ (where T = (P+N+O+U))} \rightarrow \textbf{(1)}$$

Matthews correlation coefficient (MCC)

$$MCC = \frac{(P \times N) - (O \times U)}{\sqrt{(P+U) \times (P+O) \times (N+U) \times (N+O)}} \rightarrow \textbf{(2)}$$

Sensitivity (Q_{sens})

$$Q_{sens} = \frac{P}{P+U} \rightarrow \textbf{(3)}$$

specificity (Q_{spec})

$$Q_{spec} = \frac{N}{N+O} \rightarrow \textbf{(4)}$$

Q_{Pred} (Probability of correct prediction)

$$Q_{pred} = \frac{P}{P+O} \times 100 \rightarrow \textbf{(5)}$$

where *P* and *N* refer to correctly predicted enzymes and non-enzymes, and *O* and *U* refer to over and under predictions, respectively.

ISSN 0973-2063
Bioinformation 2(3):107-112 (2007)