MATCH
Communications in Mathematical and in Computer Chemistry

MATCH Commun. Math. Comput. Chem. **69** (2013) 7-24

ISSN 0340 - 6253

Assignment of Protein Sequences to Protein Family Profiles Using Spatial Statistics

Vahid Rezaei[a], Hamid Pezeshk[b,c,*], Mehdi Sadeghi[d], Changiz Eslahchi[e]

[a] *Faculty of Mathematical Science, Tarbiat Modares University, Tehran, Iran*
[b] *School of Mathematics, Statistics and Computer Science, College of Science, University of Tehran, Iran*
[c] *Bioinformatics Research Group, School of Computer Science, Institute for Research in Fundamental Sciences (IPM), Tehran, Iran*
[d] *National Institute of Genetics Engineering and Biotechnology, Tehran-Karaj Highway, Tehran, Iran*
[e] *Faculty of Mathematical Science, Shahid Beheshti University, G. C., Tehran, Iran*

(Received September 13, 2010)

Abstract

A central problem in genomics is to determine the functions of newly discovered proteins using the information contained in their amino acid sequences. In this research we introduce a novel spatial association on a regular lattice for assignment of a protein sequence to a protein family. In our model we assume that for each residue in any position in sequence, not only the adjacent residues, but also the residues of closer homologs contain information. For this purpose we model the observation with auto correlated errors on a rectangular grid and use the information of the left, right, top and bottom residues of each amino acid in any position in a multiple sequence alignment (MSA) of the query sequence with members of each family. The spatial statistics for analyzing these observations is applied and the classification problem is solved by computing the probability of query sequence belonging to each protein family. The classification is based on the family whose MSA yields the highest probability. Using actual data, the application of spatial prediction for assignment of protein sequence to the protein profiles is proposed and the performance of the model is assessed. According to the spatial associations on a regular lattice, we use top ten profiles in the Pfam database that are very different from each other for analyzing amino acid sequences in a profile. Results show that in all cases protein sequences are assigned correctly to the corresponding protein profiles.

[*] Corresponding author
Email address: pezeshk@khayam.ut.ac.ir

1. Introduction

Proteins are the most important molecules in any living cell. They play important roles in biological phenomena and in most of the cell processes. Revealing the functions of proteins is one of the most important issues in biology. Although manual function annotation by biological experts is more reliable but it is a very time consuming procedure and sometimes an automatic approach is used. Nowadays, hundreds of complete genomes have been sequenced and a massive quantity of protein sequences has been generated. These sequences have greatest value to biology if their functions and structures can be identified. Hence a central problem in genomics is to determine the functions of newly discovered proteins using the information contained in their amino acid sequences [1]. The first approach to determine protein function is performed by doing a similarity search to find a close homologous protein with known function. Simple sequence search methods such as FASTA [2] or BLAST [3] are often used to identify close homologs of protein sequences. Although searching a sequence database using BLAST can provide satisfactory result, specially finding close homolog, but it performs poorly for homologous sequence pairs below 30-35% sequence identity [4,5,6]. One of the fundamental assumptions is that homologous protein sequences have similar functions and structures. Based on this assumption homologous sequences have been grouped into known protein families. The majority of protein sequences appear to fall into a few thousands protein families [7]. According to functional or structural relatedness, alternative family classifications have been made. Searching protein family database can produce more concise results for functional prediction of a query sequence. In the last decades, systematic methods have been developed to assign sequence to protein families [8]. Multiple alignments of a sequence family indicates different selective pressure on different residues in a functional sequence. Some positions are more conserved than others and some positions are less conserved and undergo insertions or deletions. Based on conservation pattern and position specific information in multiple sequence alignment of protein family sequences, profile methods have been introduced to search databases for homologous sequences [9,10,11]. The Hidden Markov Model (HMM) is a representation of multiple sequence alignments of protein families in term of profiles. It is commonly used in the detection of remote homologous. The Pfam [12] is a high quality set of annotated multiple alignment and pre-built profile HMM (PHMM). Given a PHMM and a protein, one can compute the probability that this protein

being generated by the PHMM using the Viterbi algorithm and infer the family that the new protein belongs to [13, 14].

In this paper we introduce a spatial association on a regular lattice for assignment of a protein sequence to a protein family. In our model we suppose that for each residue in any position in sequence, not only the adjacent residues, but also the residues of closer homologs contain information that can be used to assign a protein to a family. For this purpose we model the observation with auto correlated errors on a rectangular grid and use the information of the left, right, top and bottom residues of each amino acid in any position in a multiple sequence alignment (MSA) of query sequence with members of each family. The order in which sequences appeared in the final MSA are determined by the guide tree and hence top and bottom sequences of query sequence are the closer homologs. Spatial model for analyzing this observation is introduced and the classification problem is solved by computing the probability of query sequence belonging to each protein family. The classification is based on the family whose MSA yields the highest probability.

2. Materials and Method

2.1. Data Preparation

The Pfam [12] is a well known database of protein families. It is widely used to align new protein sequences on the known proteins of a given family or to recognize new member of a protein family. For each protein domain family in Pfam, there is a *seed alignment* which is a manually verified multiple alignment of a representative set of sequences. The Pfam database contains 11912 families (Release 24.0, October 2009). Each entry describes a domain function represented by a multiple sequence alignment and a hidden Markov model. There are two components at Pfam: Pfam-A and Pfam-B. The Pfam-A entries have high quality. As shown in table 1, we selected and used ten families from top twenty families of Pfam-A

ID	Accession	Number of sequences	
		Seed	Full
RVT 1	PF00078	155	105557
WD40	Pf00400	1843	93618
RVP	PF00033	50	76107
Cytochorm B N	PF00033	92	61850

HA TPase c	PF02518	662	57843
BPD transp 1	PF00528	81	53993
Oxidored q1	PF00361	33	52312
Pkinase	PF00069	54	51174
Adh short	PF00106	230	42056
Acetyltransf 1	PF00583	243	37490

Table 1. Top ten protein family profiles from the Pfam database

2.2. Preliminaries

In this section spatial discrete observations on a rectangular grid are introduced. The notation of this section mostly follows that of Besag [15]. The auto binomial, the auto logistic and the auto Poisson models for analyzing spatial lattice data have been considered by Besage. Jackson et al [16] described different methods for estimation of spatial parameters. The Bayesian auto logistic models have been developed by Chen et al [17]. Also, Yan et al [18] introduced spatial stochastic volatility for lattice data. In this work, it is assumed that the sites are arranged as a regular lattice, and each site has one of the c + 1 observations (multinomial data) denoted by z = 0,1,…,c. On a regular lattice the neighbors of the site with coordinates (r,s) can be denoted by {(r - 1, s), (r + 1, s),(r, s - 1),(r, s + 1)}. This can be considered as the first order Markov chain.

Besage [15] defined the following formula for computing the probability of observations P(z(s)) on a regular lattice with multinomial data:

$$p(z(s)) \propto \exp\Big(\sum_{r=1}^{c} m_r u_r + \sum_{r=1}^{c}\sum_{s=1}^{c} n_s v_s\Big) \qquad (1)$$

in which m_r is the number of sites with the rth observation, n_s is the number of pairs of neighboring sites with the observation **s**, and $\boldsymbol{u_r}$ and $\mathbf{v_s}$ are parameters. For estimating the parameters, Besage expresses (1) in the conditional form

$$p(z(s_{r,s}) = j \,/ \, othersites) = \frac{\exp(u_j + v_j n_{jrs})}{1 + \sum_{i=1}^{c} \exp(u_i + v_i n_{irs})}, \; j = 0,1,2,...,c \qquad (2)$$

where n_{jrs} is the number of sites with *jth* observations neighboring site *(r,s)*. Due to large number of parameters in this model, the pseudo-likelihood function (the product of full conditional function) that is based on conditional distribution has been used

$$p(z(s)) = \prod_r \prod_s \prod_j p(z(s_r) = j \mid others ites) = \prod_r \prod_s \prod_j \frac{\exp(u_j + v_j n_{jr})}{1 + \sum_{i=1}^{c} \exp(u_i + v_i n_{ir})} \quad (3)$$

In this paper, using equation (3) and coding method, the parameters have been estimated. Coding method is the parameter estimation method that was introduced by Besage [19]. In this method, based on the first order Markov chain model, the variables associated with the x sites, given the observed values at all other sites, are mutually independent (figure 1) and equation (3) is reduced based on the product of conditional form of x sites.

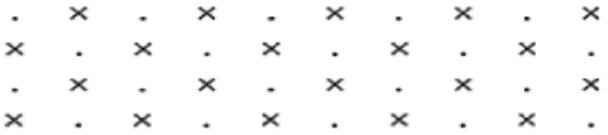

Figure 1. Coding pattern for the first- order scheme

2.3 .The application of spatial statistics in profiles of proteins

In this section the application of spatial prediction for assignment of protein sequence to the protein profiles is proposed. The performance of this model in assigning sequences to families of proteins is computed. According to the spatial associations on a regular lattice, we used top ten profiles in the Pfam database that are very different from each other for analyzing amino acid sequences in a profile.

Based on ten profiles in table 1 and using coding method in equation (3), the neighbor-dependent coefficient for each sites of a protein in a multiple alignment has been accurately estimated. In other words, v_is and u_is that represent the neighbor-dependent coefficients have been estimated. Because there are 20 different amino acids in each profile, we have 20 u_i and 20 v_i parameters used in equation (3). Since the maximization of 40 parameters are very difficult and time consuming, amino acids are classified into two classes based on their hydropathy index according to Kyte and Dolitte [20]. The hydrophilic amino acids (R, N, D, E, Q,G, H, K, P, S, T, W, Y) are classified as group1 and the hydrophilic amino acids (A, C, I, L, M, F, V) are classified as group2. By this classification only $2u_i$ and $2v_i$ parameters are used in equation (3). The parameters u_3 and v_3 indicate the gap. For example if we have a site in a multiple alignment with different amino acids neighboring such as below:

G

A A A

K

Then

$$p(z(s_r) = 1/othersites) = \frac{\exp(u_2 + v_2 * 1)}{1 + \exp(u_1 + v_1 * 1) + \exp(u_2 + v_2 * 1) + \exp(u_3)}$$

(single letter codes of A, G and K are 2,1,1 respectively)

For computing the validity of the spatial prediction for the assignment of protein sequences in protein profiles, we have performed a procedure using the following steps:

1- One sequence from each profile is removed (100 sequences that removed from 10 profiles are shown in appendix (table A1). Multiple sequence alignment (MSA) was done with sequences of each family. To perform the MSA we use CLUSTALW program that is widely used [21]. The order of sequences appeared in the MSA determined by the guide tree. Hence, the top and the bottom sequences of each sequence are it's closer homologs.

2- For each column in a profile we compute the Entropy and for each column that has entropy less than the median is removed. According to Kyte-Dollite classification, the target attribute can take either one of 2 different values or the gap. Hence the entropy of a sequence, S, is defined by

3- $Entropy(S) = \sum_{i=1}^{3} -p_i \log_2^{p_i}$

4- Where p_i is the proportion of S belonging to class i

RVT1	WD40	RVP	Cytochorm B N
>KRIT1_HUMAN/320353	>Q7NUV6_CHRVO/13181	>ARCB_SHIFL/528640	>Q35057_MARPO/544811
>Q17583_CAEEL/307339	>O28892_ARCFU/11211	>SLN1_YEAST/1090 1207	>P71276_ECOLX/62269
>Q83730_9POXV/250282	>Q2N9I4_ERYLH/5222	>BARA_ECOLI/670782	>Q46666_ECOLX/34241
>V034_FOWPV/76105	>Q73XF0_MYCPA/3235	>BVGS_BORBR/9751092	>Q47526_ECOLX/33239
>Q24241_DROME/20	>B2UEV9_RALPJ/16	>Q57N82_SALCH/81	>RT67_ECOLX/48

0232	207	21	262
>ANK1_HUMAN/172	>Q3ABV8_CARHZ/5	>PATA_ANASP/263	>RT16_MYXXA/1
204	189	375	86407
>PHO81_YEAST/5065	>Q8KEB1_CHLTE/8	>ALGB_PSEAE/1112	>Q17003_ANOGA/
38	218	1	605832
>ZDH17_HUMAN/89	>CYBH_ECOLI/1122		>O16587_CAEEL/
121	0	>YFHA_ECOLI/8118	780970
>AKR2_YEAST/8311	>A4BH07_9GAMM/6	>NTRX_AZOC5/511	>Q9YGS2_FUGRU
5	165	7	/573830
	>A5G519_GEOUR/7		>RTJK_DROFU/50
>AVO2_YEAST/3971	204	>DCTD_RHILE/7117	2757
HA TPase c	BPD transp 1	Oxidorder q1	Pkinase
>Q9KIE0_STRHY/403	>NUOL_ECOLI/1334	>Y1409_SYNY3/801	>Q9ZBG1_STRCO
64192	06	18	/128331
>Q9EWA2_9ACTO/13	>NU5M_CHLRE/132	>PRP17_HUMAN/54	>SAPB_HAEIN/93
711538	391	0579	316
>Q93NX7_9ACTO/13	>NU5M_ALBCO/963	>LE14B_PRUAR/415	>Q9HLG4_THEAC
861553	55	453	/81265
>Q9L4X2_9ACTO/499	>NU5M_ASCSU/963	>LE14B_LITER/4174	>Q9WYD8_THEM
75164	53	55	A/94282
>Q9X993_9ACTO/463	>NU5M_TRYBB/130	>Q9ZT99_ARATH/4	>Q97VF6_SULSO/
14794	400	95533	101289
>Q93NW6_9ACTO/10	>NU4M_ASPAM/130	>SWD3_YEAST/277	>YEJE_ECOLI/158
47810645	401	313	341
>Q93NW6_9ACTO/86	>NQO13_PARDE/13	>YZLL_CAEEL/3523	>Q49553_MYCHO
958862	0412	90	/234419
>Q93NW6_9ACTO/51	>NU4M_ASCSU/873	>Q93785_CAEEL/27	>Q98QS7_MYCPU
995366	46	592797	/195382
>Q93NX8_9ACTO/14	>NU4M_ARBLI/1113	>Q24593_DROME/32	>O86691_STRCO/
201587	83	833321	402574
>ERYA2_SACER/314	>NU4M_ANOGA/10	>O94042_CANAL/40	>Q987J0_RHILO/1
13306	6374	8446	00280
Adh short	Acetyltransf 1		
>O67458_AQUAE/621	>Q8UMQ4_9HIV1/39		
45	2461		
>Y1207_METJA/1282	>Q9ID52_9HIV1/238		
11	307		
>P74651_SYNY3/4112	>Q9IN30_9HIV1/408		
3	477		
>O07326_STRCO/641	>Q9IXF5_9HIV1/791		
60	48		
>YXBD_BACSU/4511	>Q99B96_9HIV1/238		
9	307		
>O05719_BACCE/481	>Q8Q7B8_9HIV1/33		
31	7406		
>Q55655_SYNY3/481	>Q87622_SIVCZ/404		
20	473		
>P73419_SYNY3/5012	>Q87108_SIVSA/425		

1	494
>O05911_MYCTU/48	>Q8AFL5_SIVCZ/14
144	6214
>YCF52_PORPU/8616	>O90273_SIVCZ/421
0	492

Table A1. 100 test sequences that are removed from 10 profiles

5- Using equation (3) the parameters v_i and u_i are estimated. The number of parameters in the conditional form, (conditional maximum likelihood estimates of the unknown parameters) can be obtained by fminunc(.) function in Matlab Software. The Fminunc(.) function can find the minimum of unconstrained multi-variable function. The Fminunc(.) attempts to find a minimum of a scalar function of several variables, starting at an initial estimate. This is generally referred to as unconstrained nonlinear optimization. x = fminunc(-fun,x_0) starts from the point x_0 and attempts to find a local maximum x of the function described in fun(.). We generate x_0 (u_1, u_2, u_3 and v_1,v_2,v_3) 1000 times to find the maximum of the conditional likelihood function.

6- Finally, each sequence that is removed from a profile in step 1, is added to all profiles. Based on parameter estimation and conditional form in equation (2), we compute the probability of the sequence being in the profile. We use the logarithm of probability divided by the length of each profile to assign each sequence to a protein family.

3. Result and Discussion

We first estimate the neighbor-dependent coefficient for each site using coding method in equation (3). Table 2 shows the estimated parameters; v_i and u_i for each of the ten profiles. Since we classify amino acids into two groups (hydrophilics and hydrophobics) only two u_i and two v_i parameters are calculated and the third parameter is reserved for the gap position in each column of multiple sequence alignment.

parameter	Protein Profiles				
	Cytochrom B N	Pkinase	RVP	RVT 1	adh short
U1	19.0508	14.9401	16.1354	16.1176	14.618
U2	18.0369	13.704	14.4894	14.8062	13.7032
U3	14.6757	11.0899	11.6607	12.0359	11.1094
V1	-0.1701	-0.0613	-0.142	-0.1153	-0.0321

0232	207	21	262
>ANK1_HUMAN/172	>Q3ABV8_CARHZ/5	>PATA_ANASP/263	>RT16_MYXXA/1
204	189	375	86407
>PHO81_YEAST/5065	>Q8KEB1_CHLTE/8	>ALGB_PSEAE/1112	>Q17003_ANOGA/
38	218	1	605832
>ZDH17_HUMAN/89	>CYBH_ECOLI/1122		>O16587_CAEEL/
121	0	>YFHA_ECOLI/8118	780970
>AKR2_YEAST/8311	>A4BH07_9GAMM/6	>NTRX_AZOC5/511	>Q9YGS2_FUGRU
5	165	7	/573830
	>A5G519_GEOUR/7		>RTJK_DROFU/50
>AVO2_YEAST/3971	204	>DCTD_RHILE/7117	2757
HA TPase_c	BPD transp 1	Oxidorder q1	Pkinase
>Q9KIE0_STRHY/403	>NUOL_ECOLI/1334	>Y1409_SYNY3/801	>Q9ZBG1_STRCO
64192	06	18	/128331
>Q9EWA2_9ACTO/13	>NU5M_CHLRE/132	>PRP17_HUMAN/54	>SAPB_HAEIN/93
711538	391	0579	316
>Q93NX7_9ACTO/13	>NU5M_ALBCO/963	>LE14B_PRUAR/415	>Q9HLG4_THEAC
861553	55	453	/81265
>Q9L4X2_9ACTO/499	>NU5M_ASCSU/963	>LE14B_LITER/4174	>Q9WYD8_THEM
75164	53	55	A/94282
>Q9X993_9ACTO/463	>NU5M_TRYBB/130	>Q9ZT99_ARATH/4	>Q97VF6_SULSO/
14794	400	95533	101289
>Q93NW6_9ACTO/10	>NU4M_ASPAM/130	>SWD3_YEAST/277	>YEJE_ECOLI/158
47810645	401	313	341
>Q93NW6_9ACTO/86	>NQO13_PARDE/13	>YZLL_CAEEL/3523	>Q49553_MYCHO
958862	0412	90	/234419
>Q93NW6_9ACTO/51	>NU4M_ASCSU/873	>Q93785_CAEEL/27	>Q98QS7_MYCPU
995366	46	592797	/195382
>Q93NX8_9ACTO/14	>NU4M_ARBLI/1113	>Q24593_DROME/32	>O86691_STRCO/
201587	83	833321	402574
>ERYA2_SACER/314	>NU4M_ANOGA/10	>O94042_CANAL/40	>Q987J0_RHILO/1
13306	6374	8446	00280
Adh short	Acetyltransf 1		
>O67458_AQUAE/621	>Q8UMQ4_9HIV1/39		
45	2461		
>Y1207_METJA/1282	>Q9ID52_9HIV1/238		
11	307		
>P74651_SYNY3/4112	>Q9IN30_9HIV1/408		
3	477		
>O07326_STRCO/641	>Q9IXF5_9HIV1/791		
60	48		
>YXBD_BACSU/4511	>Q99B96_9HIV1/238		
9	307		
>O05719_BACCE/481	>Q8Q7B8_9HIV1/33		
31	7406		
>Q55655_SYNY3/481	>Q87622_SIVCZ/404		
20	473		
>P73419_SYNY3/5012	>Q87108_SIVSA/425		

Table A2. The logarithm of probability divided by the length of each profile for 100 test sequences (top,bottom, right and left sides)

	RVT1	WD40	RVP	Cytochorm B N	HA TPase c	BPD transp l	Oxidorder q1	Pkinase	Adh short	Acetyltransf l
RVT1										
>KRIT1_HUMAN/320353	-0.055	-0.089	-0.086	-0.093	-0.078	-0.089	-0.076	-0.094	-0.079	-0.092
>Q17583_CAEEL/307339	-0.051	-0.088	-0.086	-0.089	-0.078	-0.089	-0.077	-0.091	-0.079	-0.092
>Q83730_9POXV/250282	-0.045	-0.089	-0.086	-0.089	-0.076	-0.087	-0.077	-0.091	-0.079	-0.092
>V034_FOWPV/76105	-0.058	-0.097	-0.09	-0.090	-0.078	-0.087	-0.079	-0.090	-0.078	-0.090
>Q24241_DROME/200232	-0.056	-0.086	-0.085	-0.089	-0.075	-0.084	-0.080	-0.090	-0.076	-0.090
>ANK1_HUMAN/172204	-0.046	-0.085	-0.084	-0.088	-0.076	-0.085	-0.080	-0.088	-0.081	-0.090
>PHO81_YEAST/506538	-0.055	-0.084	-0.099	-0.092	-0.078	-0.088	-0.082	-0.087	-0.081	-0.088
>ZDH17_HUMAN/89121	-0.046	-0.093	-0.088	-0.093	-0.074	-0.089	-0.084	-0.088	-0.080	-0.087
>AKR2_YEAST/83115	-0.051	-0.088	-0.085	-0.09	-0.078	-0.086	-0.083	-0.089	-0.081	-0.089
>AVO2_YEAST/3971	-0.049	-0.089	-0.092	-0.09	-0.076	-0.088	-0.080	-0.09	-0.080	-0.09
WD40										
>Q7NUV6_CHRVO/13181	-0.08	-0.058	-0.064	-0.090	-0.088	-0.079	-0.07	-0.068	-0.078	-0.093
>O28892_ARCFU/11211	-0.082	-0.061	-0.067	-0.09	-0.087	-0.076	-0.071	-0.068	-0.075	-0.093
>Q2N9I4_ERYLH/5222	-0.086	-0.058	-0.068	-0.088	-0.085	-0.079	-0.072	-0.069	-0.079	-0.090
>Q73XF0_MYCPA/3235	-0.088	-0.06	-0.069	-0.088	-0.087	-0.077	-0.069	-0.069	-0.078	-0.089
>B2UEV9_RALPJ/16207	-0.083	-0.062	-0.07	-0.089	-0.087	-0.079	-0.065	-0.072	-0.08	-0.089
>Q3ABV8_CARHZ/5189	-0.087	-0.062	-0.065	-0.091	-0.088	-0.079	-0.066	-0.071	-0.079	-0.088
>Q8KEB1_CHLTE/8218	-0.08	-0.063	-0.064	-0.090	-0.086	-0.078	-0.068	-0.07	-0.079	-0.088
>CYBH_ECOLI/11220	-0.086	-0.059	-0.066	-0.089	-0.088	-0.077	-0.067	-0.067	-0.077	-0.090
>A4BH07_9GAMM/6165	-0.086	-0.058	-0.066	-0.089	-0.089	-0.078	-0.066	-0.066	-0.081	-0.089
>A5G519_GEOUR/7204	-0.084	-0.059	-0.07	-0.088	-0.086	-0.078	-0.068	-0.07	-0.079	-0.09
RVP										
>ARCB_SHIFL/528640	-0.089	-0.089	-0.058	-0.088	-0.098	-0.078	-0.068	-0.091	-0.08	-0.089
>SLN1_YEAST/10901207	-0.089	-0.090	-0.053	-0.088	-0.099	-0.076	-0.068	-0.091	-0.083	-0.087
>BARA_ECOLI/670782	-0.088	-0.089	-0.055	-0.087	-0.099	-0.081	-0.071	-0.089	-0.082	-0.087
>BVGS_BORBR/9751092	-0.091	-0.088	-0.058	-0.086	-0.097	-0.082	-0.071	-0.089	-0.080	-0.088
>Q57N82_SALCH/8121	-0.090	-0.088	-0.056	-0.088	-0.097	-0.076	-0.070	-0.089	-0.079	-0.090
>PATA_ANASP/263375	-0.091	-0.089	-0.058	-0.090	-0.098	-0.076	-0.070	-0.092	-0.080	-0.086
>ALGB_PSEAE/11121	-0.088	-0.086	-0.053	-0.091	-0.099	-0.082	-0.067	-0.089	-0.078	-0.088
>YFHA_ECOLI/8118	-0.089	-0.085	-0.053	-0.090	-0.098	-0.08	-0.068	-0.092	-0.083	-0.085
>NTRX_AZOC5/5117	-	-0.091	-0.057	-0.089	-	-	-0.068	-0.091	-	-0.085

	RVT1	WD40	RVP	Cytochorm B N	HA TPase c	BPD transp 1	Oxidorder q1	Pkinase	Adh short	Acetyltransf 1
>DCTD_RHILE/7117	0.089	-0.091	-0.057	-0.088	-0.096	-0.08	-0.069	-0.088	-0.080	-0.075
Cytochorm B N										
>Q35057_MARPO/544811	-0.072	-0.08	-0.07	-0.069	-0.085	-0.073	-0.070	-0.078	-0.072	-0.082
>P71276_ECOLX/62269	-0.072	-0.080	-0.072	-0.07	-0.086	-0.073	-0.070	-0.077	-0.072	-0.081
>Q46666_ECOLX/34241	-0.075	-0.082	-0.072	-0.066	-0.082	-0.075	-0.071	-0.077	-0.067	-0.082
>Q47526_ECOLX/33239	-0.072	-0.080	-0.071	-0.071	-0.083	-0.075	-0.070	-0.075	-0.071	-0.079
>RT67_ECOLX/48262	-0.076	-0.078	-0.072	-0.072	-0.082	-0.072	-0.071	-0.072	-0.071	-0.079
>RT16_MYXXA/186407	-0.075	-0.078	-0.072	-0.069	-0.085	-0.076	-0.069	-0.071	-0.070	-0.077
>Q17003_ANOGA/605832	-0.072	-0.079	-0.072	-0.069	-0.084	-0.071	-0.071	-0.075	-0.070	-0.078
>O16587_CAEEL/780970	-0.074	-0.081	-0.069	-0.067	-0.084	-0.07	-0.069	-0.074	-0.068	-0.08
>Q9YGS2_FUGRU/573830	-0.071	-0.080	-0.071	-0.07	-0.088	-0.073	-0.07	-0.073	-0.07	-0.081
>RTJK_DROFU/502757	-0.072	-0.078	-0.072	-0.068	-0.082	-0.073	-0.071	-0.075	-0.069	-0.82
HA TPase c										
>Q9KIE0_STRHY/40364192	-0.069	-0.114	-0.082	-0.078	-0.065	-0.089	-0.076	-0.092	-0.09	-0.078
>Q9EWA2_9ACTO/13711538	-0.071	-0.099	-0.083	-0.079	-0.066	-0.089	-0.075	-0.091	-0.091	-0.076
>Q93NX7_9ACTO/13861553	-0.070	-0.0981	-0.079	-0.077	-0.067	-0.09	-0.074	-0.09	-0.091	-0.075
>Q9L4X2_9ACTO/49975164	-0.068	-0.0976	-0.079	-0.082	-0.065	-0.09	-0.076	-0.092	-0.087	-0.072
>Q9X993_9ACTO/46314794	-0.070	-0.102	-0.08	-0.081	-0.067	-0.089	-0.077	-0.088	-0.089	-0.073
>Q93NW6_9ACTO/1047810645	-0.068	-0.102	-0.081	-0.082	-0.068	-0.091	-0.073	-0.089	-0.089	-0.079
>Q93NW6_9ACTO/86958862	-0.070	-0.108	-0.078	-0.079	-0.068	-0.088	-0.076	-0.093	-0.088	-0.076
>Q93NW6_9ACTO/51995366	-0.071	-0.106	-0.083	-0.078	-0.069	-0.086	-0.075	-0.087	-0.09	-0.077
>Q93NX8_9ACTO/14201587	-0.072	-0.097	-0.082	-0.079	-0.07	-0.087	-0.074	-0.088	-0.091	-0.075
>ERYA2_SACER/31413306	-0.07	-0.11	-0.08	-0.08	-0.066	-0.092	-0.078	-0.093	-0.09	-0.76
BPD transp 1										
>NUOL_ECOLI/133406	-0.099	-0.098	-0.09	-0.093	-0.088	-0.065	-0.08	-0.089	-0.08	-0.073
>NU5M_CHLRE/132391	-0.072	-0.099	-0.099	-0.091	-0.088	-0.063	-0.079	-0.089	-0.08	-0.074
>NU5M_ALBCO/96355	-0.072	-0.0981	-0.0981	-0.09	-0.087	-0.063	-0.078	-0.88	-0.081	-0.076
>NU5M_ASCSU/96353	-0.073	-0.097	-0.0976	-0.09	-0.087	-0.066	-0.082	-0.087	-0.082	-0.076
>NU5M_TRYBB/130400	-0.068	-0.10	-0.102	-0.088	-0.089	-0.062	-0.081	-0.088	-0.078	-0.078
>NU4M_ASPAM/130401	-0.071	-0.102	-0.102	-0.087	-0.091	-0.06	-0.08	-0.09	-0.079	-0.076
>NQO13_PARDE/130412	-0.068	-0.102	-0.108	-0.088	-0.088	-0.063	-0.079	-0.09	-0.08	-0.075
>NU4M_ASCSU/87346	-0.069	-0.11	-0.106	-0.089	-0.086	-0.062	-0.077	-0.091	-0.083	-0.074
>NU4M_ARBLI/111383	-0.072	-0.098	-0.097	-0.09	-0.087	-0.062	-0.079	-0.092	-0.084	-0.075
>NU4M_ANOGA/106374	-	-0.1	-0.11	-0.091	-	-	-0.082	-0.089	-	-0.071

	RVT1	WD40	RVP	Cytochorm B N	HA TPase c	BPD transp l	Oxidorder q1	Pkinase	Adh short	Acetyltransf l
	0.069				0.092	0.061			0.081	
Oxidorder q1										
>Y1409_SYNY3/80118	-0.083	-0.093	-0.099	-0.09	-0.091	-0.077	-0.065	-0.097	-0.092	-0.089
>PRP17_HUMAN/540579	-0.081	-0.092	-0.12	-0.09	-0.087	-0.077	-0.064	-0.098	-0.095	-0.090
>LE14B_PRUAR/415453	-0.085	-0.092	-0.119	-0.089	-0.086	-0.081	-0.064	-0.098	-0.092	-0.087
>LE14B_LITER/417455	-0.083	-0.094	-0.101	-0.091	-0.089	-0.08	-0.064	-0.094	-0.088	-0.092
>Q9ZT99_ARATH/495533	-0.088	-0.092	-0.11	-0.089	-0.088	-0.079	-0.063	-0.093	-0.089	-0.094
>SWD3_YEAST/277313	-0.082	-0.093	-0.111	-0.088	-0.087	-0.079	-0.064	-0.097	-0.094	-0.092
>YZLL_CAEEL/352390	-0.084	-0.093	-0.111	-0.09	-0.091	-0.076	-0.063	-0.094	-0.093	-0.092
>Q93785_CAEEL/27592797	-0.081	-0.096	-0.106	-0.092	-0.9	-0.078	-0.065	-0.093	-0.091	-0.093
>Q24593_DROME/32833321	-0.082	-0.091	-0.115	-0.089	-0.086	-0.078	-0.064	-0.09	-0.090	-0.094
>O94042_CANAL/408446	-0.086	-0.093	-0.098	-0.09	-0.088	-0.077	-0.066	-0.091	-0.088	-0.093
Pkinase										
>Q9ZBG1_STRCO/128331	-0.09	-0.084	-0.075	-0.077	-0.091	-0.079	-0.073	-0.07	-0.079	-0.081
>SAPB_HAEIN/93316	-0.09	-0.08	-0.072	-0.077	-0.084	-0.078	-0.074	-0.069	-0.078	-0.081
>Q9HLG4_THEAC/81265	-0.091	-0.085	-0.069	-0.074	-0.092	-0.081	-0.074	-0.067	-0.078	-0.077
>Q9WYD8_THEMA/94282	-0.088	-0.085	-0.07	-0.073	-0.083	-0.073	-0.076	-0.068	-0.077	-0.08
>Q97VF6_SULSO/101289	-0.087	-0.084	-0.072	-0.073	-0.091	-0.074	-0.072	-0.07	-0.079	-0.081
>YEJE_ECOLI/158341	-0.089	-0.086	-0.071	-0.079	-0.086	-0.078	-0.075	-0.069	-0.081	-0.075
>Q49553_MYCHO/234419	-0.087	-0.084	-0.068	-0.072	-0.087	-0.079	-0.073	-0.071	-0.08	-0.078
>Q98QS7_MYCPU/195382	-0.088	-0.083	-0.07	-0.077	-0.089	-0.081	-0.077	-0.072	-0.077	-0.079
>O86691_STRCO/402574	-0.09	-0.085	-0.069	-0.075	-0.085	-0.076	-0.077	-0.067	-0.079	-0.078
>Q987J0_RHILO/100280	-0.089	-0.085	-0.071	-0.074	-0.084	-0.075	-0.074	-0.07	-0.08	-0.076
Adh short										
>O67458_AQUAE/62145	-0.097	-0.076	-0.088	-0.087	-0.077	-0.099	-0.073	-0.091	-0.072	-0.077
>Y1207_METJA/128211	-0.09	-0.075	-0.091	-0.087	-0.081	-0.096	-0.081	-0.090	-0.071	-0.081
>P74651_SYNY3/41123	-0.091	-0.078	-0.09	-0.089	-0.079	-0.18	-0.076	-0.092	-0.07	-0.076
>O07326_STRCO/64160	-0.09	-0.076	-0.092	-0.089	-0.078	-0.199	-0.079	-0.093	-0.068	-0.078
>YXBD_BACSU/45119	-0.097	-0.075	-0.089	-0.09	-0.08	-0.097	-0.079	-0.092	-0.069	-0.077
>O05719_BACCE/48131	-0.091	-0.077	-0.089	-0.086	-0.078	-0.1	-0.077	-0.092	-0.069	-0.081
>Q55655_SYNY3/48120	-0.092	-0.078	-0.091	-0.085	-0.078	-0.099	-0.078	-0.092	-0.067	-0.076

	RVT1	WD40	RVP	Cytochorm B N	HA TPase c	BPD transp l	Oxidorder q1	Pkinase	Adh short	Acetyltransf l
>P73419_SYNY3/50121	-0.09	-0.079	-0.09	-0.89	-0.076	-0.167	-0.079	-0.095	-0.071	-0.082
>O05911_MYCTU/48144	-0.092	-0.077	-0.088	-0.084	-0.075	-0.098	-0.078	-0.093	-0.072	-0.078
>YCF52_PORPU/86160	-0.095	-0.076	-0.092	-0.084	-0.078	-0.1	-0.08	-0.094	-0.067	-0.077
Acetyltransf 1										
>Q8UMQ4_9HIV1/392461	-0.074	-0.088	-0.096	-0.085	-0.077	-0.076	0.077	-0.072	-0.089	-0.063
>Q9ID52_9HIV1/238307	-0.076	-0.089	-0.097	-0.084	-0.08	-0.079	-0.074	-0.068	-0.09	-0.067
>Q9IN30_9HIV1/408477	-0.073	-0.088	-0.097	-0.084	-0.076	-0.075	-0.074	-0.073	-0.085	-0.062
>Q9IXF5_9HIV1/79148	-0.071	-0.091	-0.097	-0.089	-0.079	-0.075	-0.075	-0.07	-0.087	-0.065
>Q99B96_9HIV1/238307	-0.07	-0.087	-0.099	-0.05	-0.074	-0.078	-0.076	-0.071	-0.089	-0.064
>Q8Q7B8_9HIV1/337406	-0.075	-0.091	-0.096	-0.086	-0.075	-0.076	-0.078	-0.073	-0.086	-0.064
>Q87622_SIVCZ/404473	-0.076	-0.09	-0.099	-0.089	-0.077	-0.075	-0.079	-0.068	-0.09	-0.062
>Q87108_SIVSA/425494	-0.075	-0.092	-0.099	-0.87	-0.074	-0.073	-0.079	-0.067	-0.085	-0.062
>Q8AFL5_SIVCZ/146214	-0.072	-0.091	-0.095	-0.083	-0.072	-0.08	-0.078	-0.07	-0.086	-0.065
>O90273_SIVCZ/421492	-0.072	-0.091	-0.099	-0.09	-0.073	-0.075	-0.075	-0.069	-0.091	-0.066

Table A3: logarithm of probability divided by length of each profile for 100 test sequences (right and left hand sides)

	RVT1	WD40	RVP	Cytochorm B N	HA TPase c	BPD transp l	Oxidorder q1	Pkinase	Adh short	Acetyltransf l
RVT1										
>KRIT1_HUMAN/320353	-0.074	-0.075	-0.083	-0.095	-0.08	-0.077	-0.086	-0.078	-0.09	-0.081
>Q17583_CAEEL/307339	-0.082	-0.074	-0.082	-0.098	-0.08	-0.079	-0.089	-0.08	-0.089	-0.083
>Q83730_9POXV/250282	-0.076	-0.073	-0.080	-0.095	-0.082	-0.076	-0.09	-0.084	-0.089	-0.08
>V034_FOWPV/76105	-0.086	-0.068	-0.079	-0.095	-0.078	-0.079	-0.086	-0.082	-0.091	-0.077
>Q24241_DROME/200232	-0.073	-0.073	-0.077	-0.096	-0.077	-0.076	-0.086	-0.081	-0.085	-0.078
>ANK1_HUMAN/172204	-0.073	-0.074	-0.078	-0.092	-0.081	-0.077	-0.083	-0.08	-0.088	-0.078
>PHO81_YEAST/506538	-0.085	-0.07	-0.076	-0.097	-0.082	-0.079	-0.086	-0.082	-0.089	-0.084
>ZDH17_HUMAN/89121	-0.081	-0.076	-0.078	-0.094	-0.081	-0.078	-0.084	-0.080	-0.087	-0.08
>AKR2_YEAST/83115	-0.078	-0.069	-0.083	-0.095	-0.078	-0.076	-0.083	-0.077	-0.088	-0.078
>AVO2_YEAST/3971	-0.08	-0.070	-0.083	-0.098	-0.078	-0.075	-0.09	-0.082	-0.087	-0.08
WD40										
>Q7NUV6_CHRVO/13181	-0.084	-0.072	-0.061	-0.084	-0.077	-0.079	-0.07	-0.074	-0.075	-0.076
>O28892_ARCFU/11211	-0.082	-0.074	-0.059	-0.082	-0.074	-0.079	-0.067	-0.077	-0.076	-0.076
>Q2N9I4_ERYLH/5222	-0.086	-0.075	-0.065	-0.086	-0.075	-0.081	-0.062	-0.074	-0.077	-0.078

	RVT1	WD40	RVP	Cytochorm B N	HA TPase c	BPD transp 1	Oxidorder q1	Pkinase	Adh short	Acetyltransf 1
>Q73XF0_MYCPA/3235	-0.084	-0.074	-0.055	-0.084	-0.073	-0.079	-0.062	-0.078	-0.08	-0.078
>B2UEV9_RALPJ/16207	-0.081	-0.07	-0.07	-0.081	-0.073	-0.078	-0.062	-0.078	-0.076	-0.081
>Q3ABV8_CARHZ/5189	-0.083	-0.075	-0.062	-0.083	-0.072	-0.08	-0.067	-0.081	-0.075	-0.081
>Q8KEB1_CHLTE/8218	-0.08	-0.068	-0.059	-0.08	-0.073	-0.081	-0.063	-0.077	-0.075	-0.078
>CYBH_ECOLI/11220	-0.086	-0.068	-0.068	-0.086	-0.071	-0.08	-0.062	-0.077	-0.075	-0.081
>A4BH07_9GAMM/6165	-0.081	-0.071	-0.06	-0.081	-0.077	-0.078	-0.066	-0.082	-0.079	-0.081
>A5G519_GEOUR/7204	-0.084	-0.074	-0.061	-0.084	-0.075	-0.079	-0.061	-0.08	-0.075	-0.08
RVP										
>ARCB_SHIFL/528640	-0.07	-0.085	-0.057	-0.09	-0.084	-0.088	-0.087	-0.082	-0.08	-0.083
>SLN1_YEAST/10901207	-0.07	-0.082	-0.063	-0.09	-0.084	-0.088	-0.087	-0.081	-0.08	-0.083
>BARA_ECOLI/670782	-0.07	-0.083	-0.062	-0.091	-0.082	-0.09	-0.86	-0.082	0.081	-0.082
>BVGS_BORBR/9751092	-0.068	-0.084	-0.063	-0.092	-0.082	-0.091	-0.086	-0.084	-0.082	-0.078
>Q57N82_SALCH/8121	-0.072	-0.078	-0.062	-0.089	-0.079	-0.08	-0.089	-0.08	-0.079	-0.078
>PATA_ANASP/263375	-0.068	-0.079	-0.062	-0.088	-0.078	-0.091	-0.09	-0.08	-0.079	-0.08
>ALGB_PSEAE/11121	-0.07	-0.081	-0.057	-0.087	-0.08	-0.088	-0.087	-0.082	-0.08	-0.082
>YFHA_ECOLI/8118	-0.069	-0.085	-0.063	-0.093	-0.089	-0.089	-0.089	-0.083	-0.082	-0.081
>NTRX_AZOC5/5117	-0.068	-0.079	-0.061	-0.091	-0.087	-0.088	-0.087	-0.079	-0.079	-0.078
>DCTD_RHILE/7117	-0.069	-0.081	-0.058	-0.09	-0.08	-0.9	-0.88	-0.081	-0.08	-0.079
Cytochorm B N										
>Q35057_MARPO/544811	-0.07	-0.089	-0.095	-0.073	-0.087	-0.072	-0.079	-0.078	-0.076	-0.083
>P71276_ECOLX/62269	-0.074	-0.093	-0.095	-0.074	-0.086	-0.073	-0.079	-0.079	-0.076	-0.081
>Q46666_ECOLX/34241	-0.072	-0.094	-0.097	-0.071	-0.083	-0.075	-0.078	-0.079	-0.075	-0.083
>Q47526_ECOLX/33239	-0.069	-0.089	-0.097	-0.069	-0.087	-0.069	-0.081	-0.075	-0.074	-0.079
>RT67_ECOLX/48262	-0.069	-0.091	-0.092	-0.075	-0.089	-0.073	-0.08	-0.073	-0.079	-0.08
>RT16_MYXXA/186407	-0.07	-0.095	-0.091	-0.073	-0.085	-0.078	-0.077	-0.072	-0.078	-0.078
>Q17003_ANOGA/605832	-0.071	-0.093	-0.09	-0.07	-0.083	-0.076	-0.078	-0.076	-0.074	-0.077
>O16587_CAEEL/780970	-0.068	-0.09	-0.095	-0.069	-0.087	-0.067	-0.078	-0.073	-0.072	-0.077
>Q9YGS2_FUGRU/573830	-0.069	-0.089	-0.092	-0.073	-0.085	-0.072	-0.079	-0.074	-0.073	-0.078
>RTJK_DROFU/502757	-0.07	-0.091	-0.093	-0.074	-0.085	-0.07	-0.081	-0.076	-0.072	-0.077
-HA TPase c										
>Q9KIE0_STRHY/40364192	-0.071	-0.075	-0.082	-0.082	-0.069	-0.073	-0.079	-0.092	-0.084	-0.075
>Q9EWA2_9ACTO/13711538	-0.074	-0.074	-0.083	-0.081	-0.068	-0.07	-0.079	-0.089	-0.085	-0.073
>Q93NX7_9ACTO/13861553	-0.073	-0.074	-0.08	-0.081	-0.07	-0.072	-0.075	-0.089	-0.086	-0.076
>Q9L4X2_9ACTO/49975164	-0.076	-0.075	-0.081	-0.08	-0.069	-0.069	-0.077	-0.091	-0.085	-0.071

	RVT1	WD40	RVP	Cytochorm B N	HA TPase c	BPD transp 1	Oxidorder q1	Pkinase	Adh short	Acetyltransf 1
>Q9X993_9ACTO/46314794	-0.070	-0.073	-0.08	-0.082	-0.071	-0.07	-0.078	-0.088	-0.088	-0.074
>Q93NW6_9ACTO/1047810645	-0.072	-0.075	-0.081	-0.081	-0.072	-0.071	-0.08	-0.091	-0.084	-0.073
>Q93NW6_9ACTO/86958862	-0.076	-0.073	-0.079	-0.081	-0.067	-0.075	-0.081	-0.092	-0.088	-0.074
>Q93NW6_9ACTO/51995366	-0.078	-0.072	-0.084	-0.08	-0.072	-0.073	-0.077	-0.088	-0.084	-0.075
>Q93NX8_9ACTO/14201587	-0.072	-0.073	-0.078	-0.081	-0.07	-0.07	-0.079	-0.089	-0.085	-0.074
>ERYA2_SACER/31413306	-0.073	-0.075	-0.078	-0.078	-0.072	-0.069	-0.081	-0.09	-0.086	-0.75
BPD transp 1										
>NUOL_ECOLI/133406	-0.08	-0.079	-0.069	-0.07	-0.078	-0.077	-0.067	-0.089	-0.084	-0.073
>NU5M_CHLRE/132391	-0.081	-0.078	-0.069	-0.071	-0.08	-0.078	-0.07	-0.088	-0.084	-0.073
>NU5M_ALBCO/96355	-0.082	-0.077	-0.07	-0.072	-0.79	-0.074	-0.067	-0.89	-0.083	-0.072
>NU5M_ASCSU/96353	-0.083	-0.076	-0.07	-0.073	-0.08	-0.08	-0.073	-0.09	-0.081	-0.075
>NU5M_TRYBB/130400	-0.084	-0.076	-0.068	-0.069	-0.081	-0.066	-0.066	-0.089	-0.087	-0.075
>NU4M_ASPAM/130401	-0.078	-0.078	-0.069	-0.071	-0.082	-0.075	-0.071	-0.091	-0.086	-0.076
>NQO13_PARDE/130412	-0.078	-0.079	-0.071	-0.073	-0.077	-0.071	-0.072	-0.092	-0.082	-0.074
>NU4M_ASCSU/87346	-0.079	-0.079	-0.072	-0.068	-0.078	-0.073	-0.068	-0.087	-0.084	-0.073
>NU4M_ARBLI/111383	-0.077	-0.08	-0.069	-0.069	-0.079	-0.079	-0.067	-0.091	-0.087	-0.077
>NU4M_ANOGA/106374	-0.078	-0.077	-0.067	-0.07	-0.081	-0.08	-0.067	-0.09	-0.083	-0.073
Oxidorder q1										
>Y1409_SYNY3/80118	-0.091	-0.084	-0.092	-0.098	-0.06	-0.078	-0.083	-0.098	-0.076	-0.073
>PRP17_HUMAN/540579	-0.092	-0.085	-0.091	-0.098	-0.061	-0.078	-0.079	-0.099	-0.075	-0.073
>LE14B_PRUAR/415453	-0.093	-0.086	-0.092	-0.099	-0.065	-0.077	-0.075	-0.099	-0.078	-0.074
>LE14B_LITER/417455	-0.089	-0.083	-0.093	-0.098	-0.07	-0.076	-0.07	-0.097	-0.076	-0.075
>Q9ZT99_ARATH/495533	-0.088	-0.085	-0.092	-0.1	-0.05	-0.077	-0.079	-0.098	-0.075	-0.076
>SWD3_YEAST/277313	-0.089	-0.085	-0.091	-0.1	-0.063	-0.078	-0.076	-0.097	-0.074	-0.073
>YZLL_CAEEL/352390	-0.091	-0.084	-0.091	-0.097	-0.062	-0.08	-0.075	-0.096	-0.075	-0.074
>Q93785_CAEEL/27592797	-0.09	-0.086	-0.093	-0.099	-0.06	-0.08	-0.079	-0.099	-0.073	-0.075
>Q24593_DROME/32833321	-0.093	-0.087	-0.094	-0.097	-0.063	-0.079	-0.074	-0.095	-0.075	-0.076
>O94042_CANAL/408446	-0.09	-0.083	-0.095	-0.097	-0.069	-0.078	-0.068	-0.099	-0.077	-0.078
Pkinase										
>Q9ZBG1_STRCO/128331	-0.083	-0.075	-0.073	-0.087	-0.091	-0.073	-0.086	-0.071	-0.079	-0.087
>SAPB_HAEIN/93316	-0.084	-0.076	-0.073	-0.088	-0.084	-0.075	-0.085	-0.073	-0.079	-0.086
>Q9HLG4_THEAC/81265	-0.084	-0.078	-0.072	-0.089	-0.092	-0.073	-0.087	-0.072	-0.08	-0.087
>Q9WYD8_THEMA/94282	-0.08	-0.074	-0.071	-0.09	-0.083	-0.072	-0.088	-0.071	-0.078	-0.09

	RVT1	WD40	RVP	Cytochrom B N	HA TPase c	BPD transp l	Oxidorder q1	Pkinase	Adh short	Acetyltransf l
>Q97VF6_SULSO/101289	-0.085	-0.075	-0.075	-0.084	-0.091	-0.074	-0.085	-0.074	-0.079	-0.088
>YEJE_ECOLI/158341	-0.079	-0.076	-0.07	-0.082	-0.086	-0.075	-0.084	-0.072	-0.081	-0.086
>Q49553_MYCHO/234419	-0.084	-0.078	-0.074	-0.087	-0.087	-0.078	-0.086	-0.073	-0.082	-0.085
>Q98QS7_MYCPU/195382	-0.082	-0.075	-0.073	-0.088	-0.089	-0.079	-0.089	-0.069	-0.078	-0.084
>O86691_STRCO/402574	-0.08	-0.073	-0.072	-0.089	-0.085	-0.076	-0.083	-0.069	-0.078	-0.086
>Q987J0_RHILO/100280	-0.084	-0.074	-0.073	-0.089	-0.084	-0.074	-0.086	-0.073	-0.079	-0.089

Adh short

>O67458_AQUAE/62145	-0.074	-0.08	-0.072	-0.08	-0.078	-0.081	-0.087	-0.083	-0.07	-0.087
>Y1207_METJA/128211	-0.072	-0.081	-0.074	-0.084	-0.076	-0.082	-0.086	-0.084	-0.073	-0.085
>P74651_SYNY3/41123	-0.073	-0.079	-0.073	-0.082	-0.077	-0.08	-0.089	-0.094	-0.072	-0.086
>O07326_STRCO/64160	-0.074	-0.083	-0.075	-0.08	-0.0775	-0.083	-0.087	-0.08	-0.071	-0.088
>YXBD_BACSU/45119	-0.072	-0.076	-0.076	-0.081	-0.08	-0.0084	-0.09	-0.081	-0.072	-0.089
>O05719_BACCE/48131	-0.074	-0.08	-0.077	-0.083	-0.079	-0.082	-0.088	-0.086	-0.073	-0.09
>Q55655_SYNY3/48120	-0.077	-0.079	-0.07	-0.083	-0.081	-0.081	-0.091	-0.087	-0.075	-0.085
>P73419_SYNY3/50121	-0.074	-0.083	-0.072	-0.82	-0.075	-0.078	-0.085	-0.085	-0.071	-0.086
>O05911_MYCTU/48144	-0.073	-0.084	-0.074	-0.086	-0.076	-0.08	-0.087	-0.084	-0.073	-0.087
>YCF52_PORPU/86160	-0.074	-0.076	-0.076	-0.086	-0.077	-0.082	-0.086	-0.087	-0.072	-0.089

Acetyltransf 1

>Q8UMQ4_9HIV1/392461	-0.08	-0.076	-0.089	-0.082	-0.07	-0.071	-0.078	-0.076	-0.085	-0.069
>Q9ID52_9HIV1/238307	-0.083	-0.075	-0.087	-0.08	-0.07	-0.069	-0.076	-0.077	-0.082	-0.07
>Q9IN30_9HIV1/408477	-0.082	-0.073	-0.085	-0.082	-0.07	-0.069	-0.084	-0.072	-0.083	-0.074
>Q9IXF5_9HIV1/79148	-0.079	-0.075	-0.089	-0.085	-0.075	-0.074	-0.078	-0.076	-0.088	-0.071
>Q99B96_9HIV1/238307	-0.078	-0.07	-0.084	-0.085	-0.07	-0.075	-0.083	-0.07	-0.083	-0.069
>Q8Q7B8_9HIV1/337406	-0.08	-0.075	-0.088	-0.086	-0.074	-0.074	-0.081	-0.072	-0.087	-0.072
>Q87622_SIVCZ/404473	-0.083	-0.075	-0.085	-0.079	-0.071	-0.076	-0.076	-0.073	-0.088	-0.07
>Q87108_SIVSA/425494	-0.082	-0.075	-0.088	-0.08	-0.072	-0.071	-0.084	-0.072	-0.087	-0.069
>Q8AFL5_SIVCZ/146214	-0.081	-0.074	-0.085	-0.081	-0.073	-0.074	-0.083	-0.072	-0.088	-0.074
>O90273_SIVCZ/421492	-0.08	-0.074	-0.084	-0.081	-0.069	-0.075	-0.082	-0.073	-0.088	-0.074

Acknowledgments. Vahid Rezaei is grateful to the Department of Statistics at Tarbiat Modares University. Hamid Pezeshk would like to thank the Department of Research Affairs of University of Tehran. This work is in part supported by IPM (No. CS 1389-0-01).

References

[1] R.M. Karp, Mathematics challenges from genomics and molecular biology, *Notices Amer. Math. Soc.* **49** (2002) 544–553.

[2] W. R. Pearson, D. J. Lipman, Improved tools for biological sequence comparison, *Proc. Natl. Acad. Sci. USA* **85** (1988) 2444–2448.

[3] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, Basic local alignment search tool, *J. Mol. Biol.* **215** (1990) 403–410.

[4] W. R. Pearson, Comparison of methods for searching protein sequence databases, *Protein Sci.* **4** (1995) 1145–1160.

[5] P. Agrawal, D. J. States, Comparative accuracy of methods for protein sequence similarity search, *Bioinformatics* **14** (1998) 40–47.

[6] B. Rost, Twilight zone of protein sequence alignments, *Protein Engin. Design Selection* **12** (1998) 85–94.

[7] C. Chothia, One thousand families for the molecular biologist, *Nature* **357** (1992) 543–544.

[8] W. R. Pearson, M. L. Sierk, The limits of protein sequence comparison, *Curr. Opin. Struct. Biol* **15** (2005) 254–260.

[9] W. R. Taylor, Identification of protein sequence homology by consensus template alignment, *J. Mol. Biol*. **188** (1986) 233–258.

[10] G. J. Barton, Protein multiple sequence alignment and flexible pattern matching, *Methods Enzymol.* **183** (1990) 403–428.

[11] S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Res.* **25** (1997) 3389–3402.

[12] R. D. Finn, J. Mistry, J. Tate, P. Coggill, A. Heger, J. E. Pollington, O. L. Gavin, P. Gunasekaran, G. Ceric, K. Forslund, S. R. Eddy, E. L. L. Sonnhammer, A. Bateman, The Pfam protein families database, *Nucleic Acids Res.* **38** (2010) D211-D222.

[13] S. R. Eddy, Profile hidden Markov models, *Bioinformatics* **14** (1998) 755–763.

[14] A. Krogh, M. Brown, I. S. Mian, K. Sjölander, D. Haussler, Hidden Markov models in computational biology. Applications to protein modeling, *J. Mol. Biol.* **235** (1994) 1501–1531.

[15] J. Besag, Spatial interaction and the statistical analysis of lattice systems, *J. Royal Stat. Soc.* **36B** (1974) 192–236.

[16] M. C. Jackson, *Spatial data analysis for discrete data on a lattice*, Phd Thesis, University of Maryland, Department of Mathematics, 2003.

[17] F. Chen, *Bayesian modeling of multivariate spatial binary data with application to the distribution of plant species*, Phd Thesis, Department of Statistics, Florida State University, 2002.

[18] J. Yan, Spatial stochastic volatility for lattice data, *J. Agricult. Biol. Environ. Stat.* **12** (2007) 25–40.

[19] J. Besage, C. Kooperberg, On conditional and intrinsic auto regression, *Biometrika* **82** (1995) 733–746.

[20] J. Kyte, R. Doolittle, A simple method for displaying the hydropathic character of a protein, *J. Mol. Biol.* **157** (1982) 105–132.

[21] M. A. Larkin, G. Blackshields, N. P. Brown, R. Chenna, P. A. McGettigan, H. McWilliam, F. Valentin, I. M. Wallace, A. Wilm, R. Lopez, J. D. Thompson, T. J. Gibson, D. G. Higgins, Clustal W and Clustal X version *2*.0, *Bioinformatics* **23** (2007) 2947–2948.