Predicting helix-helix interactions from residue contacts in membrane proteins

Allan Lo[1,2], Yi-Yuan Chiu[3], Einar Andreas Rødland[4,5], Ping-Chiang Lyu[2], Ting-Yi Sung[3,*], and Wen-Lian Hsu[3,*]

[1]Bioinformatics Program, Taiwan International Graduate Program, Academia Sinica, Taipei, [2]Department of Life Sciences, National Tsing Hua University, Hsinchu, [3]Bioinformatics Lab., Institute of Information Science, Academia Sinica, Taipei, Taiwan, [4]Centre for Cancer Biomedicine, University of Oslo, NO-0027 Oslo, and [5]Norwegian Computing Center, P.O. Box 114 Blindern, NO-0314 Oslo, Norway

SUPPLEMENTARY INFORMATION

Details of statistical methods for contact propensities

Single residue and residue pair contact propensities were estimated using a beta-binomial model. Let i denote the residue type. The data then consist of values n_i and N_i for each i indicating that there are n_i contacts out of a total of N_i possible contacts. For single residues, N_i counts all residues that may form contacts, i.e., all residues within the transmembrane region. For residue pairs, N_{ij} counts all possible contact pairs made of from single contact residues on the set of interacting helices in the same protein. Thus, the single residue contact probabilities have been accounted for, and in a way that ensures that compositional differences between proteins should not produce spurious contact propensities.

For clarity, we only describe single residue contact propensities in the following discussion. Calculations of residue pair contact propensities follow the same procedure as the single residue contact propensities. We assume that n_i is drawn from a binomial distribution with probability P_i: i.e., each of the N_i possible contacts have a probability P_i of being an actual contact. This is a slight simplification, in particular for contact pairs, as contacts are, strictly speaking, not independent, yet we expect this to be at most of minor importance. One possible weakness is that in large proteins with many contact residues, N_i will be high since it counts the total number of residues, and it may be that the actual number of contacts increases in proportion to the number of possible contact pairs; however, we expect this to have at most a modest effect which should not produce a systematic bias. For contact pair computations, the numbers may be low, and for the less frequent residues, the ratio n_i/N_i might be heavily influenced by randomness. This would give rise to contact propensities that differ substantially from 1, but with very broad confidence intervals. This may be avoided by assuming that the contact

[*] To whom correspondence should be addressed.

probabilities P_i also fit a distribution. In effect, this distribution amounts to adding an *a priori* assumption on what are likely values of the P_{ij}. We use a beta distribution: this is a sufficiently general distribution to fit both the mean and variation of the P_i, and is a natural choice as it is conjugate to the binomial distribution and thus make the computations much easier.

To summarise the beta-binomial model, we assume that there are parameters M and μ of the beta distribution

$$\beta(p; M\mu, M(1-\mu)) = \frac{\Gamma(M)p^{M\mu}(1-p)^{M(1-\mu)}}{\Gamma(M\mu)\Gamma\big(M(1-\mu)\big)}$$

from which the P_i is derived. The binomial distribution of n is then given by

$$f(n; N, p) = \frac{N!}{n!\,(N-n)!} \cdot p^n(1-p)^{N-n}$$

which gives the posterior distribution of p given n as

$$\frac{\beta(p; M\mu, M(1-\mu)) \cdot f(n; N, p)}{\int_0^1 \beta(p; M\mu, M(1-\mu)) \cdot f(n; N, p)\, dp} = \beta(p; n + M\mu, N - n + M(1-\mu))$$

which makes the *a posteriori* estimate $\hat{p}$ (either expected value or maximum likelihood) of p equal

$$\hat{p} = \frac{n + M\mu}{N + M}$$

rather than the direct ratio n/N. The terms $M\mu$ and M which are added in the numerator and denominator are often referred to as “shrinkage factors”, and these may also be deduced or motivated using methods other than the beta-binomial model. In effect, we see that these shrinkage factors caused by the assumed beta distribution of the contact probabilities correspond to assuming M prior residues with μ being the portion of contacts amongst these. When N is large compared to M, the effect of these shrinkage factors is small. However, when N is small, more emphasis is placed on the shrinkage factors.

We have estimated the parameters M and μ using maximum likelihood estimators: i.e., the parameters maximizing the likelihood function

$$L(n; M, \mu) = \prod_i \int_0^1 \beta\big(p_i; M\mu, M(1-\mu)\big) \cdot f(n_i; N_i, p_i)\, dp_i$$

$$= \prod_i \frac{\Gamma(M)\Gamma(M\mu + n_i)\Gamma(M(1-\mu) + N_i - n_i)}{\Gamma(M + N_i)\Gamma(M\mu)\Gamma(M(1-\mu))}$$

which gives the probability of picking the list of values (n_i) for given parameters M and μ for arbitrary P_i. We have solved this by using a variety of Newton's method on the log-likelihood. To be more specific, we have solved

$$\sum_i \gamma'(M + N_i) - \gamma'(M) = \sum_i \gamma'(M\mu + n_i) - \gamma'(M\mu)$$

$$= \sum_i \gamma'(M(1-\mu) + N_i - n_i) - \gamma'(M(1-\mu))$$

where γ is the logarithm of the gamma function; its derivative, γ', is often referred to as the digamma function.

We have expressed propensities as the ratios $\hat{p}/\mu$ using the maximum likelihood estimates of M and μ. Using μ as the denominator is preferred over the overall ratio $\sum_i n_i \,/ \sum_i N_i$ since the overall ratio would place too much emphasis on the more frequent residues or residue pairs: the underlying contact probabilities of these are as variable as for the less frequent ones, and the only increase in emphasis should come from them being more accurately estimated. Uncertainties of the estimated probabilities are illustrated by the standard deviation of the *a posteriori* distributions. For residue contact propensities, the standard deviation is calculated as follows:

$$E[\hat{p}] = \frac{(M\mu + n_i)}{(M + N_i)}$$

$$Var[\hat{p}] = \frac{(M\mu + n_i)[M(1-\mu) + (N_i - n_i)]}{(M + N_i)^2(M + N_i + 1)}$$

With the propensity defined as $P_i = \hat{p}/\mu$, this makes

$$SD[P_i] = \frac{\sqrt{Var[\hat{p}]}}{\mu}$$

Bootstrapping

Since we do not have a large number of high-resolution membrane protein structures, the standard errors could not be directly estimated by assuming a normal distribution, i.e. by calculating $SE_{norm}=\frac{\sigma}{\sqrt{N}}$, where σ is the standard deviation and N is the sample size. Instead, we applied a bootstrapping procedure as a coarse-grained approximation to estimate the standard errors of performance measures. The bootstrap estimation may be viewed as how sensitively a score depends on a particular data set chosen. A similar bootstrapping method has also been used by Chen *et al*. (2002) to estimate the standard errors of accuracy by TM topology predictors on a high-resolution set of 36 membrane proteins. Briefly, we describe the bootstrapping procedure below:

1. Randomly sample *with replacement* from the original data to obtain a dataset having the same sample size of the original data.
2. Calculate the sample statistics of interest on the bootstrap sample.
3. Repeat steps 1 and 2 to obtain a large number (B) of bootstrap samples and corresponding sample statistics. Calculate the average of the statistic of interest from all the bootstrap samples.
4. Calculate the standard error (SE_{boot}) as follows:

$$SE_{boot} = \sqrt{\frac{1}{(B-1)}\sum_{b=1}^{B}(\hat{\theta}^{*b} - \overline{\theta})^2}$$

where $\hat{\theta}^{*b}$ is the value of the statistic of interest for each bootstrap sample, and $\overline{\theta}$ is the average of $\hat{\theta}^{*b}$ from B replicates.

Here, we choose B = 1000, which is generally considered a sufficient number for resampling. In addition, we have repeated the bootstrapping experiments with up to 5000 replicates, in steps of 1000, and obtained very similar results. The statistics of interest or performance measures used in the paper are accuracy, sensitivity, specificity, and *MCC*.

Table 1S. High-resolution membrane protein structures used in the development of TM*hit*.

PDB ID	Chain used	Resolution (Å)	TMH (Obs)[†]	Description
1c3w	a	1.55	7	Bacteriorhodopsin from *H. salinarium*
1dxr	m	2.00	5	Photosynthetic reaction center from *R.viridis*
1f88	a	2.80	6	Bovine rhodopsin
1h2s	a, b	1.93	(7, 2)[‡]	Sensory rhodopsin II from *N. pharaonis*
1jb0	a, l	2.50	(13, 3)	Crystal structure of photosystem I from *S.elongatus*
1ldf	a	2.10	6	Glycerol facilitator from *E. coli*
1m3x	l	2.55	5	Photosynthetic reaction center from *R.sphaeroides*
1nek	c	2.60	3	Succinate dehydrogenase from *E. coli*
1okc	a	2.20	6	Adp/Atp carrier from *B. taurus*
1pw4	a	3.30	12	Glycerol-3-phosphate transporter from *E.coli*
1r3j	c	1.90	2	Potassium channel kcsa-fab complex from *M. musculus*
1rh5	a	3.20	10	Preprotein translocase SecY from *M.jannaschii*
1s5l	d	3.50	6	Photosystem q(b) protein from *T. elongatus*
1xfh	a	3.50	10	Proton glutamate symport protein from *P. horikoshii*
1yce	a	2.40	2	F-type Na+-ATPase from *I. tartaricus*
1yew	b, c	2.80	(4, 4)	Particulate methane monooxygenase from *M. capsulatus*
2agv	a	2.40	10	Sarcoplasmic/endoplasmic reticulum calcium ATPase from *O. cuniculus*
2axt	a, c, z	3.00	(5, 7, 2)	Photosystem II from *T. elongatus*
2bs2	c	1.78	5	Quinol-fumarate reductase from *W. succinogenes*
2fbw	d	2.10	3	Respiratory complex II from *G. gallus*
2fyn	a	3.20	8	Cytochrome b from *R. sphaeroides*
2gfp	a	3.50	11	Multidrug transporter EmrD from *E. coli*
2gif	a	2.90	18	Acriflavine resistance protein b from *E. coli*
2hyd	a	3.00	6	Multidrug ABC transporter from *S. aureus*
2jaf	a	1.70	7	Halorhodopsin from *H. salinarium*
2jiz	g	2.30	2	ATP synthase from *B. taurus*
2nmr	a	2.10	9	Ammonia channel from *E. coli*
2nq2	a	2.40	8	Metal-chelate type ABC transporter from *H. influenzae*
2nr9	a	2.20	5	Rhomboid peptidase from *H. influenzae*
2pno	a	3.30	4	Leukotriene c4 synthase from *H. sapiens*
2qpe	a	2.90	11	Cytochrome ba3 oxidase from *T. thermophilus*
2qts	a	1.90	4	Acid-sensing ion channel from *G. gallus*
2r6g	f	2.80	7	Maltose transporter from *E. coli*
2r6g	g	2.80	4	Maltodextrin import ATP-binding protein from *E. coli*
2r9r	b	2.40	6	Voltage-gated potassium channel from *R.norvegicus*
2rh1	a	2.40	6	Human β2-adrenergic g protein-coupled receptor
2vl0	a	3.30	3	Ligand gated ion channel from *E. chrysanthemi*
2vpz	c	2.40	7	Thiosulfate reductase from *T. thermophilus*

Table 1S. High-resolution membrane protein structures used in the development of TM*hit*. (Cont'd)

PDB ID	Chain used	Resolution (Å)	TMH (Obs)[†]	Description
2yvx	a	2.40	6	Thiosulfate reductase from *T. thermophilus*
2zjs	y	3.20	10	SecYe translocon from *T. thermophilus*
3b4r	a	3.30	4	Zinc metalloprotease from *M. jannaschii*
3b9w	a	1.30	9	Ammonium transporter from *N.europaea*
3beh	a	3.10	5	Cyclic nucleotide regulated ion channel from *R. loti*
3d31	c	3.00	5	Sulfate/molybdate ABC transporter from *M. acetivorans*
3ddl	a	1.90	7	Xanthorhodopsin from *S. ruber*
3dh4	a	2.70	10	Sodium/sugar symporter from *V. parahaemolyticus*
3dhw	a	3.70	4	Methionine importer metni from *E. coli*

[†] TMH (Obs): Sum of all observed number of transmembrane helices of each chain from PDB file and parsed by STRIDE (Frishman and Argos, 1995).
[‡] The number in the parenthesis denotes the number of TMH of the respective chain in a protein.

Table 2S. High-resolution membrane proteins used in the independent test set.

PDB ID	Chain used	Resolution (Å)	TMH (Obs)[†]	TMH (Pred)[‡]	Description
1fft	a	3.50	11	15	Ubiquinol oxidase from *E. coli*
1kf6	d	2.70	3	3	Quinol-fumarate reductase from *E. coli*
1kpl	a	3.00	12	11	Chloride channel from *S. typhimurium*
1kqf	c	1.60	5	4	Formate dehydrogenase from *E. coli*
1nek	d	2.60	3	3	Succinate dehydrogenase from *E. coli*
1pv6	a	3.50	12	11	Lactose permease from *E. coli*
1q16	c	1.90	5	5	Nitrate reductase from *E. coli*
1qle	c	3.00	7	7	Cytochrome C oxidase from *P. denitrificans*
1xio	a	2.00	7	7	Anabaena sensory rhodopsin
2a65	a	1.65	9	13	Na+/Cl- symporters from *A. aeolicus*
2axt	b	3.00	6	6	Photosystem II from *T. elongatus*
2bl2	a	2.10	4	4	V-type ATPase from *E. hirae*
2fbw	c	2.10	3	3	Succinate dehydrogenase cytochrome b from *G. gallus*
2o9d	a	2.30	6	7	Aquaporin from *E. coli*

[†] TMH (Obs): Observed number of transmembrane helices from PDB file and parsed by STRIDE (Frishman and Argos, 1995).

[‡] TMH (Pred): Predicted number of transmembrane helices by SVM*top* (Lo *et al*., 2008).

Table 3S. Raw counts of contact residue and residue contact propensity.

Amino acid	**A**	**G**	**P**	**L**	**I**	**V**	**M**	**C**	**S**	**T**	**N**	**Q**	**H**	**D**	**E**	**K**	**R**	**F**	**W**	**Y**
Counts	481	304	87	479	286	349	145	49	178	188	52	63	52	29	52	30	62	242	78	142
P_i[†]	1.25	1.19	0.96	0.96	1.00	1.09	1.27	1.43	1.19	1.11	0.90	0.93	1.06	0.68	0.72	0.49	0.68	1.03	0.93	1.17
p-value[‡]	8.4e-10	9.3e-5	0.64	0.32	0.97	0.04	1.6e-4	1.5e-4	2.1e-03	0.06	0.28	0.44	0.56	5.7e-4	4.0e-4	3.8e-11	9.0e-6	0.56	0.41	0.02

[†] P_i is the calculated contact propensity of each amino acid.

[‡] The p-value for each amino acid is calculated from a binomial distribution with the *a priori* expected residue contact probability μ_r.

Table 4S. Raw counts of contact residue pairs.

Amino Acid	A	G	P	L	I	V	M	C	S	T	N	Q	H	D	E	K	R	F	W	Y
A	40	66	22	122	75	98	46	10	39	43	8	18	13	8	6	6	10	50	14	42
G		23	9	79	36	51	20	8	24	28	7	6	17	0	8	0	10	50	13	24
P			4	16	12	9	9	6	8	11	2	5	0	3	8	3	3	13	7	6
L				44	49	74	29	12	34	39	9	9	11	5	12	2	9	56	11	25
I					24	44	17	13	24	21	7	6	6	2	6	7	7	30	12	12
V						39	19	8	37	27	9	14	9	6	9	7	10	37	10	25
M							6	9	14	19	1	5	0	1	0	2	5	20	3	11
C								1	8	3	0	3	3	0	1	0	1	11	1	3
S									10	24	13	6	8	6	7	5	9	22	6	14
T										6	9	4	9	7	5	0	5	24	7	16
N											0	2	0	5	4	1	0	3	2	4
Q												1	1	2	1	1	4	8	0	5
H													1	1	1	0	1	3	2	4
D														0	2	2	3	0	0	1
E															0	2	2	5	2	1
K																0	0	2	0	2
R																	1	7	0	8
F																		10	9	9
W																			0	6
Y																				1

[†] For clarity, only the half on the right of the table is shown. The values are symmetric with respect to the diagonal.

Table 5S. Residue pair contact propensities (P_{ij}).

Amino Acid	A	G	P	L	I	V	M	C	S	T	N	Q	H	D	E	K	R	F	W	Y
A	0.88	0.94	0.99	1.06	1.08	1.13	1.11	1.00	1.02	1.01	0.94	1.05	1.04	1.04	0.92	1.00	0.96	0.96	0.96	1.11
G		0.97	0.91	1.03	0.93	0.98	0.93	1.01	0.99	1.01	0.98	0.94	1.14	0.94	1.02	0.93	1.03	1.14	1.03	1.00
P			1.03	0.89	1.00	0.89	1.02	1.08	1.01	1.04	1.00	1.04	0.98	1.04	1.10	1.04	1.01	1.03	1.06	0.98
L				0.84	0.87	0.95	0.89	1.03	0.95	0.95	0.96	0.91	0.99	0.97	1.03	0.93	0.95	0.99	0.89	0.89
I					1.05	0.93	0.92	1.11	1.01	0.92	1.00	0.94	0.99	0.97	0.98	1.06	1.00	0.95	1.04	0.90
V						1.09	0.89	1.01	1.12	0.93	0.99	1.06	1.05	1.02	1.01	1.04	1.02	0.96	0.97	1.04
M							0.98	1.09	1.02	1.06	0.94	1.00	0.95	0.98	0.94	0.99	1.02	1.03	0.96	1.00
C								1.01	1.07	0.99	0.98	1.03	1.03	0.99	1.00	0.99	1.00	1.10	0.99	1.00
S									1.06	1.13	1.15	1.02	1.06	1.07	1.06	1.05	1.07	1.03	1.00	1.05
T										0.96	1.07	0.97	1.10	1.08	1.01	0.94	1.00	1.04	1.02	1.05
N											0.99	1.00	0.98	1.08	1.05	1.01	0.97	0.94	1.00	1.00
Q												1.00	0.99	1.02	0.99	1.00	1.04	1.01	0.97	1.01
H													0.99	1.01	1.00	0.99	0.99	0.94	1.00	1.01
D														1.00	1.02	1.03	1.04	0.94	0.99	0.99
E															0.99	1.02	1.01	0.99	1.00	0.96
K																1.00	0.98	0.97	0.99	1.01
R																	0.99	1.00	0.98	1.07
F																		0.94	1.00	0.89
W																			0.97	1.01
Y																				0.93

†For clarity, only the half on the right of the table is shown. The values are symmetric with respect to the diagonal. Shaded values indicate statistically significant pairs with p-value <0.05. The p-value for each amino acid pair (not shown) is calculated from a binomial distribution with the *a priori* expected contact residue pair probability μ_p.

Table 6S. Contact pair prediction accuracy by leave-one-out cross validation on the development set.

Methods[†]	**Contact pair prediction**				**δ-analysis (|δ|=4)**	
	Accuracy		*IMP*	p-value	Accuracy	
Direct prediction						
TM*hit* $_{\text{L2 only}}$	10.2(±2.1[‡])%	182/1786	29.1	1.6e-200	32.6(±3.5)%	582/1786
Two-level model						
TM*hit*	12.9±(1.8)%	230/1786	36.9	5.9e-278	38.2(±3.6)%	683/1786

[†] Both diret and two-level models were trained and cross vaildated using observed information of helix definition from STRIDE (Frishman and Argos, 1995), topology from TOPDB (Tusnady *et al*., 2008) and observed RSA.

[‡] The standard error (SE_{boot}) estimated by bootstrapping follows the ± sign.

Table 7S. Contact pair prediction accuracy of two-level TM*hit* by grouping of TMH numbers from leave-one-out cross validation on the development set.

Number of TMHs	**Contact pair prediction**				**δ-analysis ($\|\delta\|$=4)**	
	Accuracy		*IMP*	p-value	Accuracy	
2-4 (*N*=16)[†]	12.2(±1.9[‡])%	36/295	34.9	9.1e-33	52.5(±5.0)%	155/295
5-6 (*N*=15)	8.5(±2.1)%	41/480	24.3	5.3e-38	36.0(±3.7)%	173/480
7-9 (*N*=11)	31.3(±7.1)%	131/419	89.4	6.3e-208	54.7(±7.9)%	229/419
10 and above (*N*=10)	3.7(±2.5)%	22/592	10.6	2.6e-19	21.3(±4.8)%	126/592
TOTAL (*N*=52)	12.9(±1.8)%	230/1786	36.9	5.9e-278	38.2(±3.6)%	683/1786

[†] *N* is the total number of protein chains in each group.

[‡] The standard error (SE_{boot}) estimated by bootstrapping follows the ± sign.

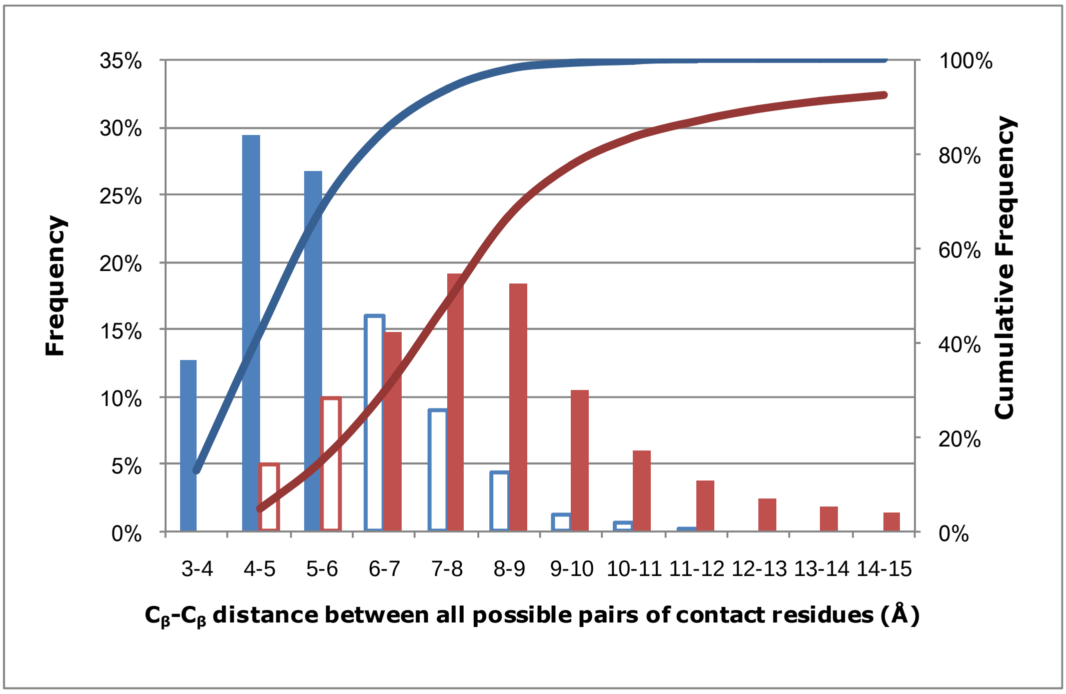

Figure 1S. The absolute and cumulative frequencies of all possible pairs of contact residues as a function of $C_ß$ distance. The residue pairs are comprised of all possible pairs of contact residues (satisfying both side-chain and backbone constraints). The $C_ß$ distance is divided into distance bins of 1Å. The blue bars (both filled and empty) indicate the frequency of residue pairs satisfying the side-chain distance criterion (interatomic distance < van der Waals radii + 0.6Å) and the fractions below 6Å indicated by blue filled bars are selected contacts. The red bars (both filled and empty) represent the fraction of residue pairs that do not satisfy the side-chain distance constraint. Here, the red filled bars with $C_ß$-$C_ß$ distance above 6Å are selected as non-contacts. The empty bars enclosed by blue and red lines represent those do not satisfy one of the constraints and hence are not selected. The cumulative frequencies of blue and red bars are shown in blue and red lines, respectively.

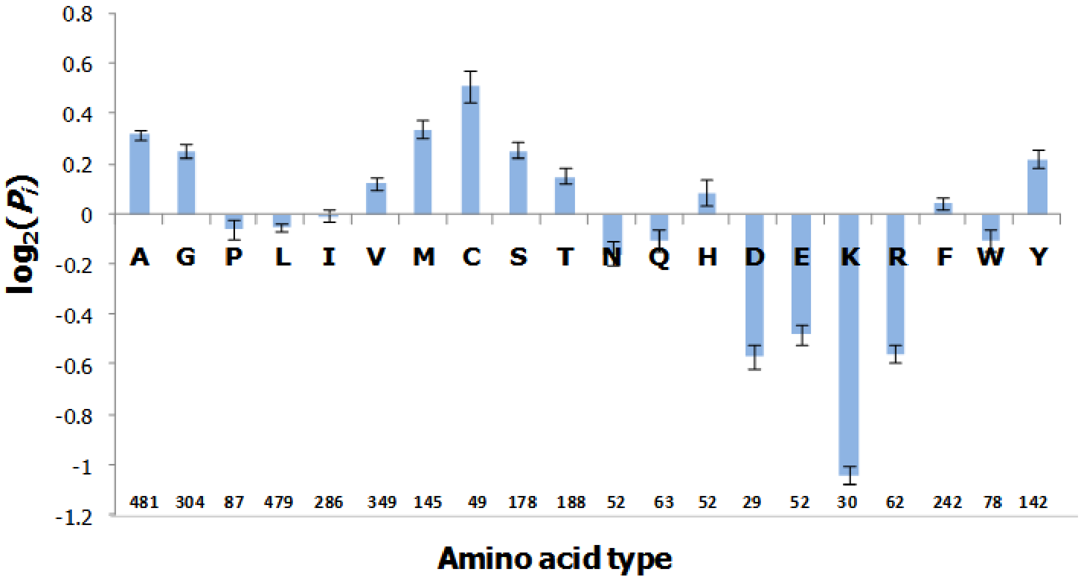

Figure 2S. Residue contact propensities shown on a $\log_2$ scale. A positive value indicates that the type of amino acid is more preferred to a contact residue than non-contact. An error bar for each propensity corresponds to the standard deviations. The number at the bottom of the horizontal axis is the count of each amino acid type occurring in residue contact pairs.

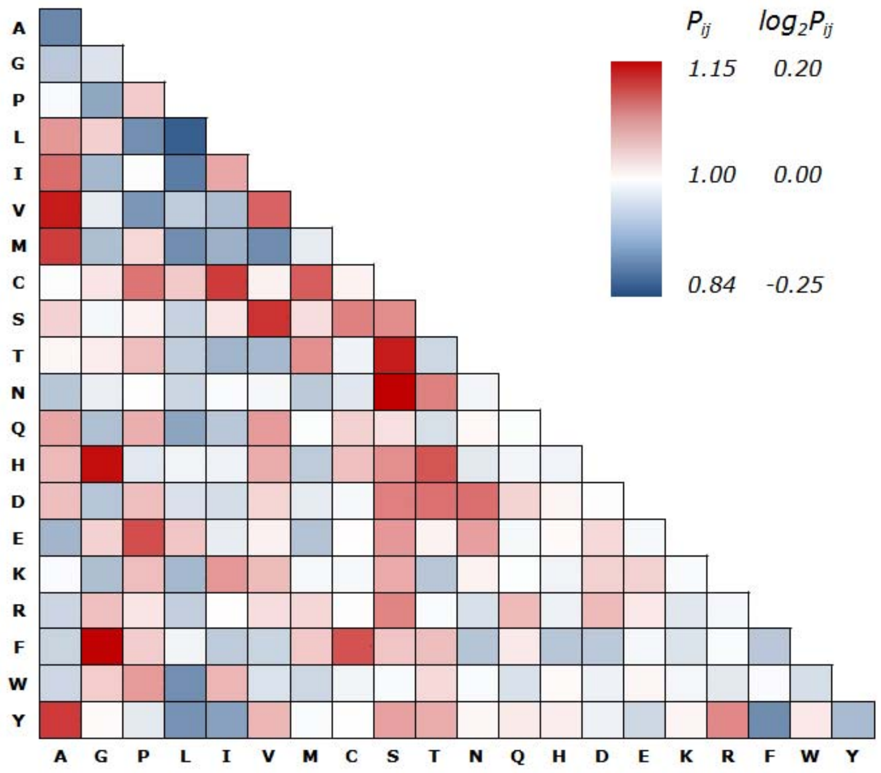

Figure 3S. Residue pair contact propensity matrix P_{ij} and its corresponding value on a $\log_2$ scale. The matrix is shown in a color-coded gradient where high propensities are represented by red; medium propensities are represented by white; and low propensities by blue.

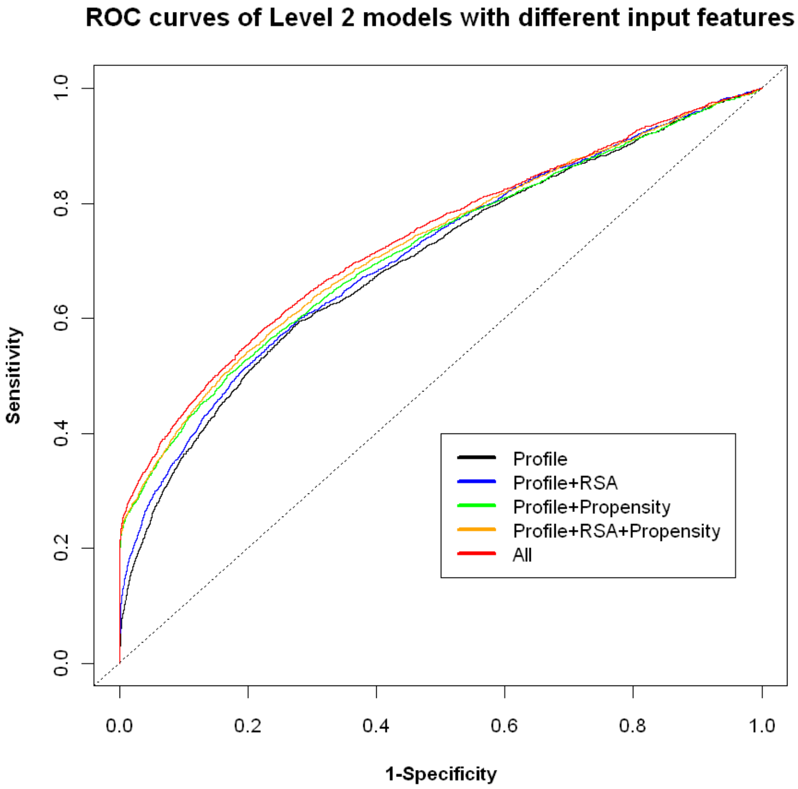

Figure 4S. Comparison of receiver operating characteristic (ROC) curves of different feature sets for Level 2 during training and LOOCV on the development set with observed information. The ROC curve of each feature set is represented by different color; i) Profile; ii) Profile+RSA; iii) Profile+ Propensity; iv) Profile+RSA+Propensity; and v) All five features: (iv+helix-heilx interaction type+helical length). The area under curve (AUC) increases for feature sets of increasing complexity. The AUC of each ROC curve from feature set i to v is 0.68, 0.70, 0.71, 0.73, and 0.75, respectively. The ROC plot was prepared using the ROCR package (Sing *et al*., 2005).

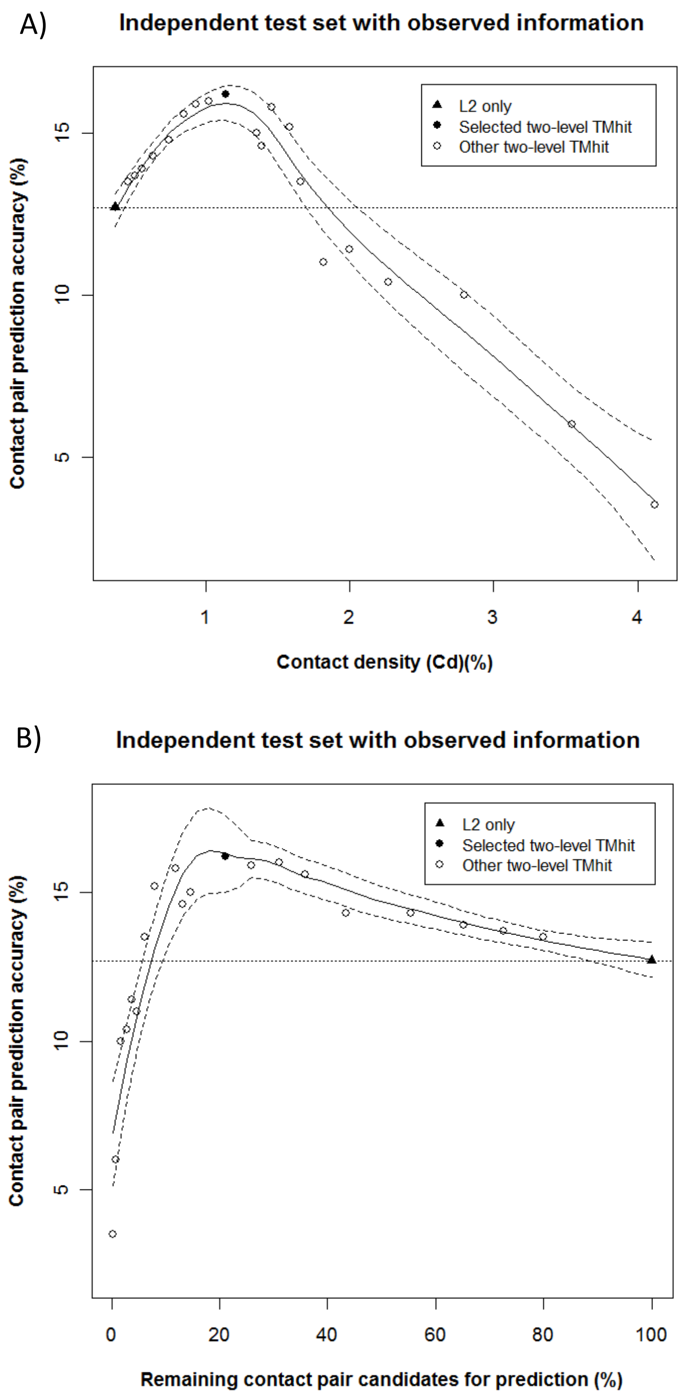

Figure 5S. Comparison of contact pair prediction accuracy as a function of contact density (*Cd*) (A) and percent remaining contact pairs candidates for prediction by Level 2 (B) by direct and two-level models on the independent test set using observed information (topology and RSA). Direct prediction (L2 only) is shown in filled triangle and its accuracy is shown in a dotted horizontal line. Two-level models are shown in filled (selected) or empty circles (others). The regression curve was estimated from all models (smoothing parameter α=0.8) using the LOCFIT package (Loader, 2004) and the dashed line indicates the confidence band at 95% confidence limits.

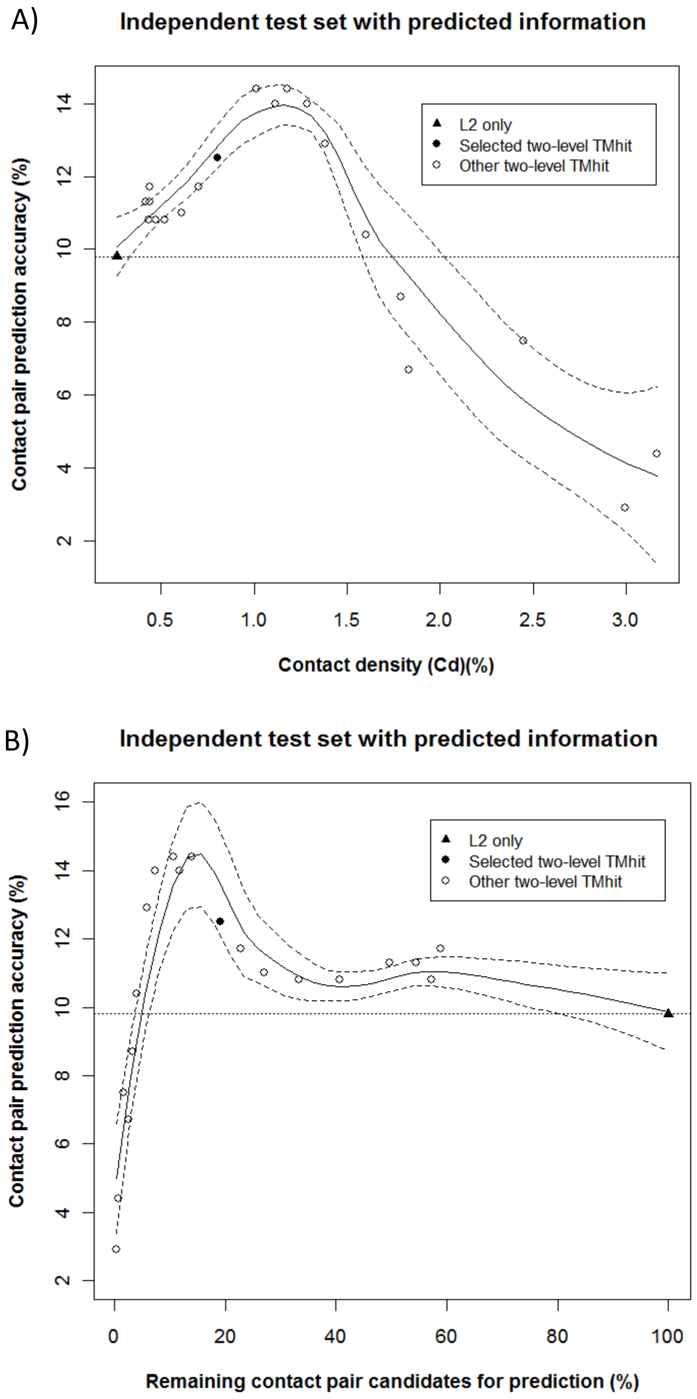

Figure 6S. Comparison of contact pair prediction accuracy as a function of contact density (*Cd*) (A) and percent remaining contact pairs candidates for prediction by Level 2 (B) by direct and two-level models on the independent test set using predicted information (topology and RSA). Direct prediction (L2 only) is shown in filled triangle and its accuracy is shown in a dotted horizontal line. Two-level models are shown in filled (selected) or empty circles (others). The regression curve was estimated from all models (smoothing parameter α=0.8) using the LOCFIT package (Loader, 2004) and the dashed line indicates the confidence band at 95% confidence limits.

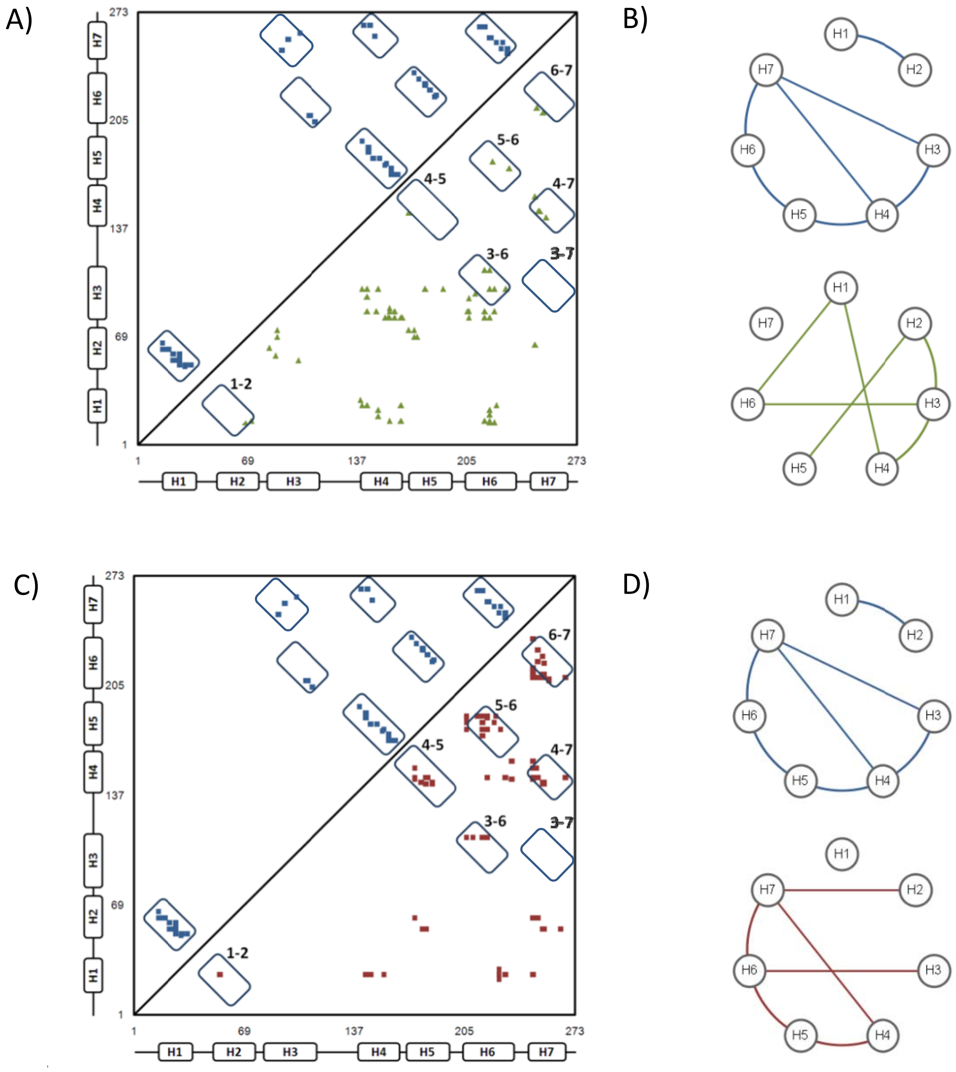

Figure 7S. Comparsion of contact maps and helix-helix interaction graphs of cytochrome c oxidase (PDB ID: 1qleC) predicted by direct and two-level TM*hit*. Observed contacts are shown in blue squares in the upper halves of the contact maps. Predicted contacts by direct method are shown in green triangles (A), and in red squares by two-level method (top $L/2$ predictions shown) (C). The observed TM helices are indicated in boxes by number along the position of the protein. Observed helix-helix interactions are labeled in boxes surrounding the contacts in the contact map. In helix-helix interaction graphs, TM helices are represented by nodes and pairwise helical interactions are represented by edges. Here, we set T=5, such that at least 5 contact pairs are required for a predicted helical interaction. The top interaction graph is the observed helix-helix interactions shown in blue edges. In (B), predicted helix-helix interactions by direct prediction are shown in green edges and in (D) by two-level are in red.

REFERENCES

Chen,C.P. *et al*. (2002) Transmembrane helix prediction revisited, *Protein Sci*., **11**, 2774-2791.

Frishman,D. and Argos,P. (1995) Knowledge-based protein secondary structure assignment, *Proteins*, **23**, 566-579.

Lo,A. *et al.* (2008) Enhanced membrane protein topology prediction using a hierarchical classification method and a new scoring function, *J Proteome Res*, **7**, 487-496.

Loader,C. (2004) Smoothing: local regression techniques. In Gentle,J. (eds), *Handbook of Computational Statistics*. Springer-Verlag, Heidelberg, pp. 540-560.

Sing,T. *et al*. (2005) ROCR: visualizing classifier performance in R, *Bioinformatics*, **21**, 3940-3941.

Tusnady,G.E. *et al*. (2008) TOPDB: topology data bank of transmembrane proteins, *Nucleic Acids Res*, ***36,*** D234-239.