ProteinPredict: A novel method for the prediction of the active site of the protein.

Pavlos Pavlidis, Michael Kokkinidis, Aikaterini Rousaki

University of Crete/Dept. of Biology & IMBB/FORTH

P.O.Box 2208, Vassilika Vouton,
GR-71409 Heraklion/Crete, Greece

ABSTRACT

Proteins and their intra and inter molecular interactions are dependent on the amino acids involved in the particular interaction. Interactions that are energetically favorable, have been evolved and optimized energetically over time. On the other hand, most energetically unfavorable interactions have been dismissed in nature. Thus, the maintenance of an unfavorable interaction in nature could be interpreted as functionally significant. Assuming that statistically frequent interactions can be interpreted as energetically favorable, we developed an algorithm to detect energetically unfavorable interactions.

Based on the study of Jha et al. (2011) we score each aminoacid of a Protein DataBank file. Jha et al. (2011) described how to summarize interactions of aminoacids in a matrix form. They demonstrated their approach on membrane proteins. Our algorithm assigns high scores to aminoacids with several statistically unfavorable interactions. The aminoacids with the highest score are considered statisistically unfavorable and thus they may have a significant functional role. We tested our algorithm on Sensory Rhodopsin II (phoborhodopsin) and detected several aminoacids that have been reported in the literature as having a significant role on the protein function. For example, Asp193 (Kitade et al. 2009; Ikeura et al. 2004), and Arg72 (Kitade et al. 2009) which are among the highest scored aminoacids have a role in the photochemistry of Sensory Rhodopsin II.

METHODS

Our methodology is based on the analysis of Jha et al. 2011 and outlined as follows:

1. We downloaded a set ***S*** of 85 helical membrane proteins from the OPM database (Lomize et al. 2006). Proteins coordinates in the OPM database have been transformed so that the x-axis is perpendicular to the membrane.
2. A connectivity matrix **A** was constructed by the set **S**. Let *i* and *j* indexing the *ith* and jth aminoacid of a protein P from S. Then,

$$A_{ij} = \begin{cases} 1 & \text{if } d(C_\alpha^i - C_\alpha^j) \le 6.5\text{Å} \quad \& \quad |i-j| > 1 \\ 0 & \text{otherwise} \end{cases}$$

where, Cia denotes the Ca atom of the ith residue.

3. The interaction matrix **M is constructed as follows:**

$$M_{ij} = -ln\frac{n_{i,j}}{g \times f_i \times f_j \times N},$$

where:
nij : number of interactions between aminoacid i and aminoacid j
g = 1 or 2. g=1 if the ith residue is the same as the same residue and 2 otherwise
fi : frequency of aminoacid i ,
N : total number of aminoacids

In practice M is constructed after aminoacids are classified into residue-contact-based environments. Only three environments have been considered here because of the small size of the dataset. All the amino acids with low and moderate degree (1-5) have been grouped together as environment I and the amino acids with degree 6 and .6 are placed in environments II and III, respectively. In this case:

$$M(i,j) = -\ln\left[\frac{n_{AI-BII}}{g \times (S_A/S) \times (S_B/S) \times (E_{I-II})}\right]$$

NAI-BII are the contacts between residues of type A that belong to environment I and residues of type B that belong to environment II.
SA is the total number of residues of type A in the dataset
SB is the total number of residues of type B in the dataset
S is the total number of residues in the dataset, and,
EI-II is the number of interactions between environment I and II.

4. Each aminoacid of a protein was scored as:

$$S_i = \sum_{j=0}^{N} A_{ij} M_{ij}$$

Analysis of Sensory Rhodopsin II

?YR	1884	124	82	2ksy.pdb	1.593411	0.398353	0.562075	*GLU*	A	GLU,LEU,PHE,GLY
?RP	2748	178	82	2ksy.pdb	1.611236	0.230177	0.811366	*LEU*	A	TYR,PRO,PHE,LEU,GLY,PRO,GLY
.EU	3118	202	82	2ksy.pdb	1.779275	0.254182	2.152487	*LYS*	A	GLY,VAL,TYR,LEU,THR,LYS,VAL
?LE	1161	77	82	2ksy.pdb	1.821595	0.364319	2.035390	*ASP*	A	ILE,ASP,THR,THR,MET
?LE	3031	197	82	2ksy.pdb	1.829161	0.365832	2.035390	*ASP*	A	ASP,VAL,ALA,TYR,LEU
?LY	2060	136	82	2ksy.pdb	1.982613	0.495653	0.875067	*ALA*	A	ALA,PHE,VAL,TYR
?SN	2523	165	82	2ksy.pdb	2.009869	0.502467	1.151376	*VAL*	A	LEU,THR,VAL,VAL
.EU	2772	179	82	2ksy.pdb	2.076856	0.296694	0.706498	*GLY*	A	PRO,PHE,ILE,GLY,PRO,GLY,VAL
?LA	1905	125	82	2ksy.pdb	2.104306	0.526076	0.851198	*GLU*	A	GLU,ARG,PHE,GLY
?RP	1137	76	82	2ksy.pdb	2.131209	0.355202	0.892396	*THR*	A	ARG,TYR,ILE,LEU,THR,THR
?HR	2944	191	82	2ksy.pdb	2.167841	0.433568	1.703938	*ASP*	A	LEU,THR,ASP,VAL,ALA
?RO	2817	182	82	2ksy.pdb	2.309800	0.577450	1.425081	*PRO*	A	GLY,VAL,ALA,PRO
?RG	2499	164	82	2ksy.pdb	2.436965	0.609241	1.223632	*THR*	A	LEU,LEU,THR,VAL
.EU	3012	196	82	2ksy.pdb	2.444193	0.488839	1.962396	*ASP*	A	ASP,VAL,VAL,TYR,LEU
?YR	1085	73	82	2ksy.pdb	2.500859	0.500172	1.625457	*ASP*	A	VAL,PRO,ASP,TRP,PHE
?LY	1718	112	82	2ksy.pdb	2.518107	0.503621	0.680153	*PHE*	A	MET,LEU,ALA,GLY,PHE
?LY	1954	128	82	2ksy.pdb	2.526902	0.505380	0.875067	*ALA*	A	TYR,ALA,LEU,GLY,ALA
?YR	739	51	82	2ksy.pdb	2.544875	0.282764	1.051173	*VAL*	A	PHE,TRP,ALA,ALA,ALA,VAL,VAL,MET,ALA
?LY	667	45	82	2ksy.pdb	2.566763	0.427794	0.930769	*GLY*	A	THR,VAL,GLY,ILE,ALA,ALA
?HR	63	5	82	2ksy.pdb	2.727018	0.681755	1.006286	*TRP*	A	LEU,PHE,TRP,ALA
?RP	130	9	82	2ksy.pdb	2.751488	0.305721	1.006286	*THR*	A	THR,THR,LEU,GLY,ALA,ILE,TYR,ALA,ALA
?HR	1199	79	82	2ksy.pdb	2.810802	0.401543	1.157033	*ASP*	A	ILE,ASP,TRP,ILE,PRO,LEU,ILE
?HR	1579	103	82	2ksy.pdb	2.840713	0.568143	1.515895	*ASN*	A	GLY,ILE,VAL,ASN,THR
.EU	2791	180	82	2ksy.pdb	2.841018	0.405860	0.811366	*TRP*	A	PHE,ILE,TRP,GLY,VAL,ALA,LEU
?RG	1061	72	82	2ksy.pdb	2.896673	0.579335	1.630986	*ASP*	A	PHE,VAL,ILE,ASP,TRP
.EU	1180	78	82	2ksy.pdb	2.952348	0.590470	1.962396	*ASP*	A	ILE,ASP,TRP,THR,PRO
?LY	855	59	82	2ksy.pdb	3.153024	0.788256	2.271398	*PRO*	A	MET,PHE,VAL,PRO
?AL	2570	168	82	2ksy.pdb	3.319110	0.474159	1.151376	*ASN*	A	VAL,ARG,ASN,LEU,LEU,TRP,ALA
?RO	2930	190	82	2ksy.pdb	3.388946	0.847236	1.694842	*ASP*	A	PRO,VAL,ASP,VAL
?LY	3277	209	82	2ksy.pdb	3.500450	0.700090	2.170864	*LYS*	A	LEU,LYS,VAL,GLY,ILE
?LE	1560	102	82	2ksy.pdb	3.608887	0.515555	1.070330	*ASN*	A	PRO,PHE,GLY,ILE,LEU,ASN,THR
?LY	630	42	82	2ksy.pdb	3.695089	0.615848	1.007528	*THR*	A	VAL,THR,LEU,SER,GLY,ILE
?LY	3250	207	82	2ksy.pdb	3.975759	0.795152	2.170864	*LYS*	A	VAL,THR,LYS,GLY,PHE
?RG	1860	123	82	2ksy.pdb	4.191149	1.047787	1.611303	*LEU*	A	GLY,ALA,LEU,PHE
?LY	2845	184	82	2ksy.pdb	4.239091	0.706515	1.149256	*PRO*	A	TRP,LEU,LEU,GLY,PRO,ALA
?YR	3066	199	82	2ksy.pdb	4.301683	0.716947	2.248759	*ASP*	A	ALA,LEU,ILE,ASP,LEU,VAL
?AL	3137	203	82	2ksy.pdb	4.638391	0.773065	3.084488	*LYS*	A	TYR,LEU,ASP,LYS,VAL,GLY
?RO	1227	81	82	2ksy.pdb	6.507624	0.929661	1.758403	*ASN*	A	LEU,THR,ILE,VAL,TYR,ILE,ASN
?RO	1047	71	82	2ksy.pdb	8.021933	1.336989	2.271398	*GLY*	A	ALA,GLY,PHE,TYR,ILE,ASP
?SN	1612	105	82	2ksy.pdb	8.036661	1.148094	1.758403	*PRO*	A	PRO,VAL,ILE,THR,VAL,VAL,MET
?SP	3106	201	82	2ksy.pdb	9.742001	1.948400	2.248759	*TYR*	A	VAL,TYR,VAL,THR,LYS
?SP	1125	75	82	2ksy.pdb	11.346650	1.620950	2.035390	*ILE*	A	ALA,PRO,ARG,TYR,ILE,LEU,THR
?SP	2974	193	82	2ksy.pdb	13.535696	1.691962	2.035390	*ILE*	A	GLY,LEU,THR,PRO,THR,ALA,LEU,ILE
.YS	3167	205	82	2ksy.pdb	13.607588	1.943941	3.084488	*VAL*	A	SER,ASP,LEU,VAL,GLY,PHE,GLY

Figure 1: Analysis of Sensory Rhodopsin II. Several residues with high score are located around the retinal molecule (Lys 205, Asp193, Arg72).

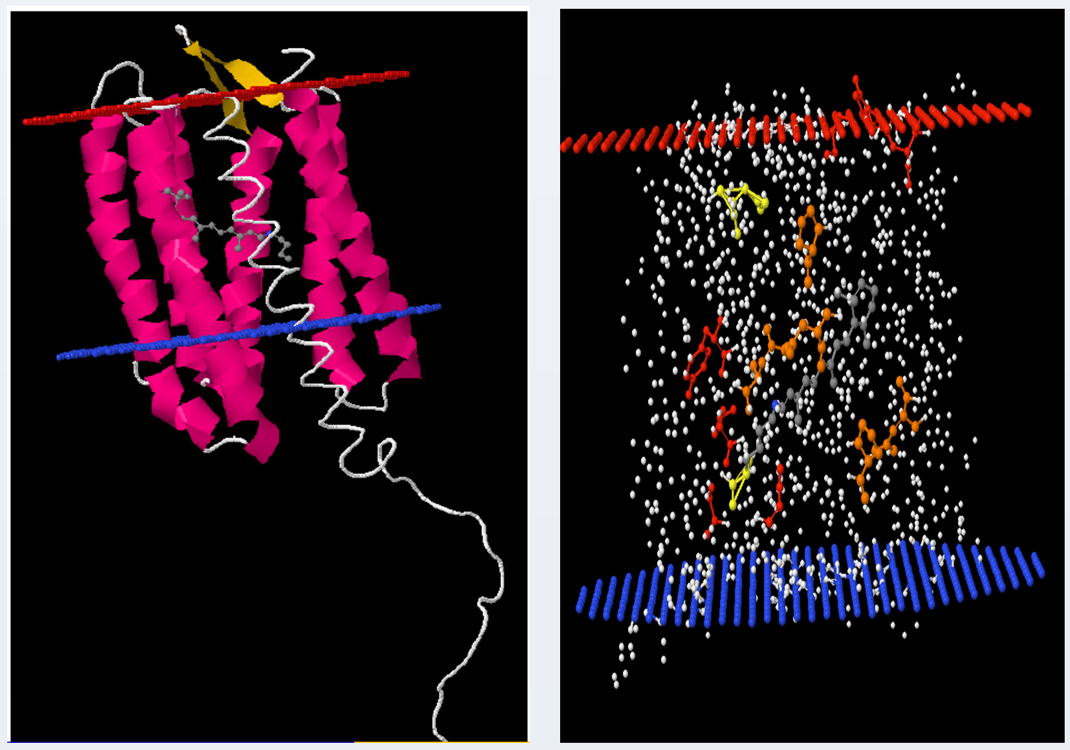

Figure 2: Analysis of Sensory Rhodopsin II. On the left panel, the 3D model of the protein is shown. On the right, residues are colored based on their score S. Red residues are those with the highest score. Several residues with high score are located around the retinal molecule (Lys 205, Asp193, Arg72).

A novel method for the prediction of the active site of the protein has been described. Sensory Rhodopsin II exemplifies the successful prediction of the retinal binding site of the protein. Lysine 205 (Gushchin et al. 2011) is the amino acid that triggers the photocycle of the protein and interacts directly with the retinal molecule. It has the highest score in Figure 1. Asp193 is mentioned to be related to the solubility of SRII (Iwamoto et al. 2002, Sudo et al. 2008) and Asp75 as a proton transfer is affected by the change of the solution pH (Jiang et al. 2010). The hits of our program are all clustered around the retinal molecule and thus determine the active site of the protein as seen in Figure 2.

References

1. Jha et al. 2011 Amino acid interaction preferences in helical membrane proteins. Protein engineering, design & selection : PEDS
2. Lomize et al. OPM database and PPM web server: resources for positioning of proteins in membranes. Nucleic Acids Res, 40(Database issue): D370-D376 (2012)
3. Gushchin et al. 2011. Active state of sensory rhodopsin II: structural determinants for signal transfer and proton pumping. J Mol Biol. 2011 Sep 30;412(4):591-600.
4. Ikeura et al. 2004. Role of Arg-72 of pharaonis Phoborhodopsin (sensory rhodopsin II) on its photochemistry. Biophys J. 2004 May;86(5):3112-20.
5. Sudo et al. 2008. A long-lived M-like state of phoborhodopsin that mimics the active state. Biophys J. 2008 Jul;95(2):753-60
6. Iwamoto et al. 2002. Role of Asp193 in chromophore-protein interaction of pharaonis phoborhodopsin (sensory rhodopsin II). Biophys J. 2002 August; 83(2): 1130–1135
7. Jiang et al. 2010. Molecular impact of the membrane potential on the regulatory mechanism of proton transfer in sensory rhodopsin II. J Am Chem Soc. 2010 Aug 11;132(31):10808-15