Motifs and Interactions in Membrane Proteins and Applications in Structure Prediction

Jie Liang

Dept. of Bioengineering
University of Illinois at Chicago

(Joint work with Ronald Jackups and Larisa Adamian)

Outline

- β-barrel membrane proteins.
  - Intrinsic residue preference for different lipid regions: Positive-outside rule
  - Sequence motifs: Ala-Tyr dichotomy, chaperon recognition site.
  - Determinants of destination of nascent peptide chains: inserted or pushed through.

  - Strand-strand interactions: Aromatic rescue.
  - Determinants of folding and assembly.
  - Structure prediction.

- α-helical membrane proteins.
  - Helix-helix interactions.
  - Helix-lipid interactions.
  - prediction.

β-barrel Membrane Proteins

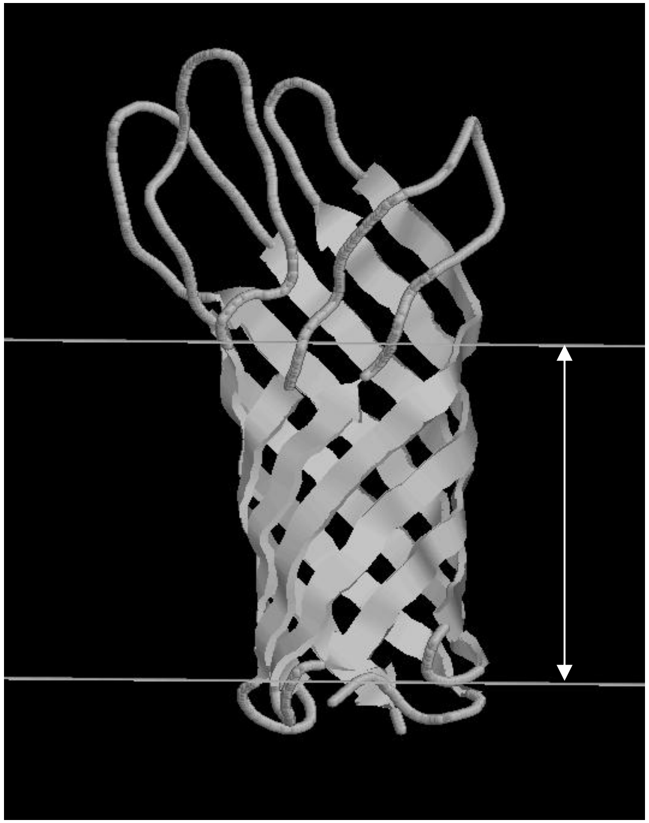
- Outer membrane of gram-negative bacteria, mitochondria, and chloroplasts.
- Acid-fast gram-positive bacteria and pore-forming exotoxin of gram-positive bacteria.
  - ~20 non-homologous structures
- Diverse function:
  - Bacteria adhesion, structural integrity, enzyme activity, colicin release, diffusion of molecules, transport of vitamine/iron, bacterial virulence, and immune surveillance.

- Medical implications:
  - bacterial infections that rely on β-barrel membrane proteins:
    - *E. coli*
    - meningitis
    - *Staphylococcus aureus*
    - anthrax
  - Targets for antibacterial drugs and vaccines.

(with Ronald Jackups)

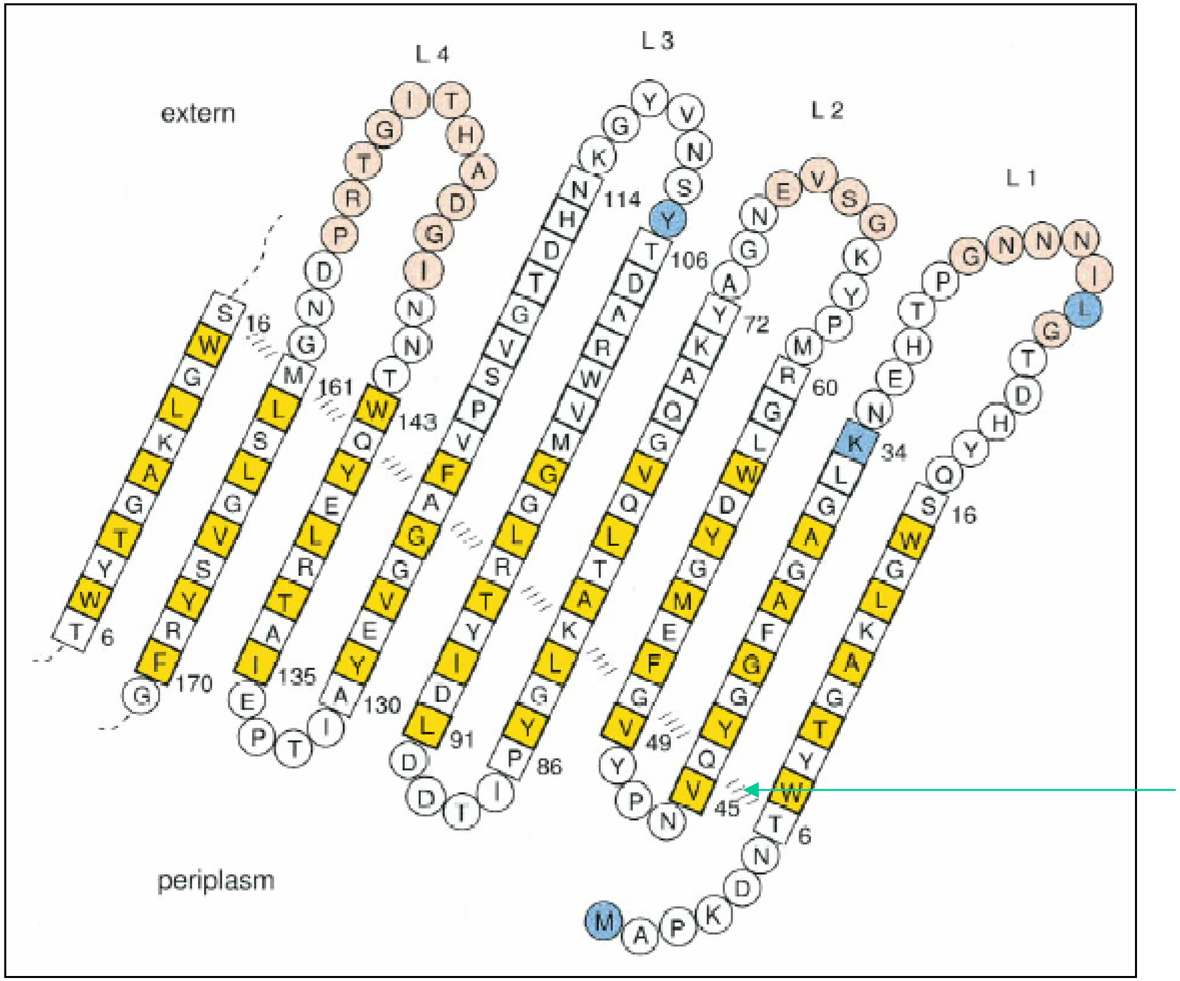
Architecture of β-Barrel Membrane Proteins. I

- Even number of strands
- Antiparallel
- Twisted and tilted into barrels
  - Right-hand twist
  - Tilt ≈ 30-60º
- Loops extend outside membrane
  - Short Pro-rich periplasmic loops
  - Long polar extracellular loops
- Monomer or oligomers



Arrow: Width of outer membrane (~27 Å)

Architecture of β-barrel Membrane Proteins. II

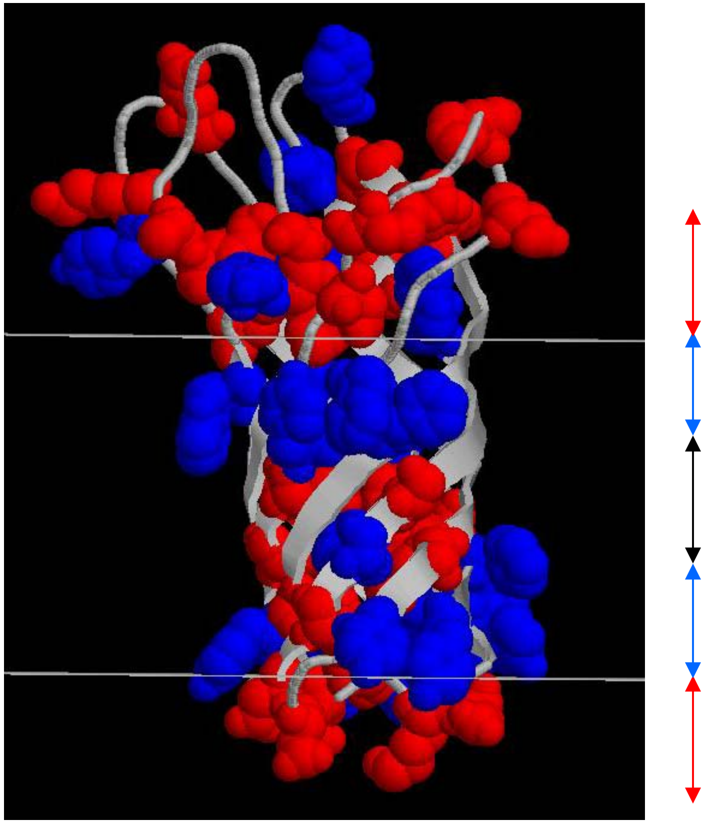


H-bonding pattern

🟨 : Outward facing

Pautsch A. and Schultz G. *Nat Struct Biol.* 1998 Nov 5(11):1013-7.

Regions of β-barrel Strands



External cap (+13.5 to +20.5)

External Headgroup (+6.5 to +13.5)

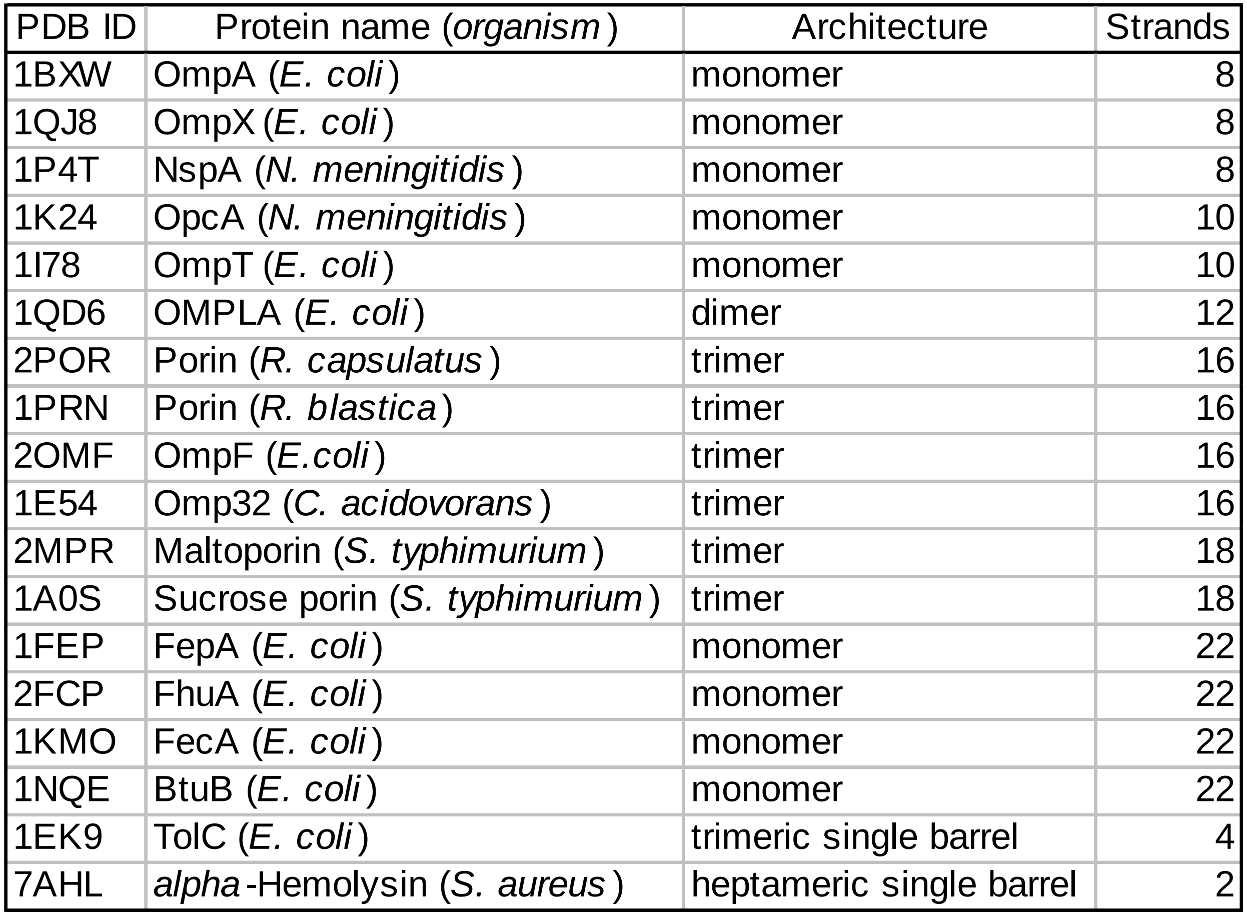
Transmembrane core (-6.5 to +6.5)

Periplasmic headgroup (-13.5 to -6.5)

Periplasmic cap (-20.5 to -13.5)

(Wimley, 2002, Prot. Sci)

Structures of 18 β-Barrel Membrane Proteins



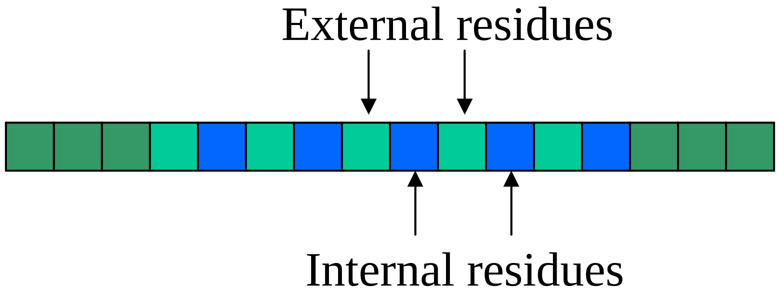
PDB ID	Protein name (*organism*)	Architecture	Strands
1BXW	OmpA (*E. coli*)	monomer	8
1QJ8	OmpX (*E. coli*)	monomer	8
1P4T	NspA (*N. meningitidis*)	monomer	8
1K24	OpcA (*N. meningitidis*)	monomer	10
1I78	OmpT (*E. coli*)	monomer	10
1QD6	OMPLA (*E. coli*)	dimer	12
2POR	Porin (*R. capsulatus*)	trimer	16
1PRN	Porin (*R. blastica*)	trimer	16
2OMF	OmpF (*E.coli*)	trimer	16
1E54	Omp32 (*C. acidovorans*)	trimer	16
2MPR	Maltoporin (*S. typhimurium*)	trimer	18
1A0S	Sucrose porin (*S. typhimurium*)	trimer	18
1FEP	FepA (*E. coli*)	monomer	22
2FCP	FhuA (*E. coli*)	monomer	22
1KMO	FecA (*E. coli*)	monomer	22
1NQE	BtuB (*E. coli*)	monomer	22
1EK9	TolC (*E. coli*)	trimeric single barrel	4
7AHL	*alpha*-Hemolysin (*S. aureus*)	heptameric single barrel	2

All proteins share less than 15% sequence identity by BLAST search.

Residue preference for different regions

(Jackups, Jr and JL, *J Mol Biol*, 2005)

- Residues are alternatingly internally and externally facing.



- Preference of residue types in different regions of TM β-barrel.

- Odds ratio $P_r\,(X)$: observed frequency $f\,(X|r)$ of residue type i against expected frequency $\mathbb{E}[f'(X|r)]$ in region R:

$$P_r(X) = \frac{f(X|r)}{\mathbb{E}[f'(X|r)]}.$$

The Null Model

- Exhaustive permutation of all residues in all regions.
- Each permutation occurs with equal probability.
  - Assign regions based on their new positions

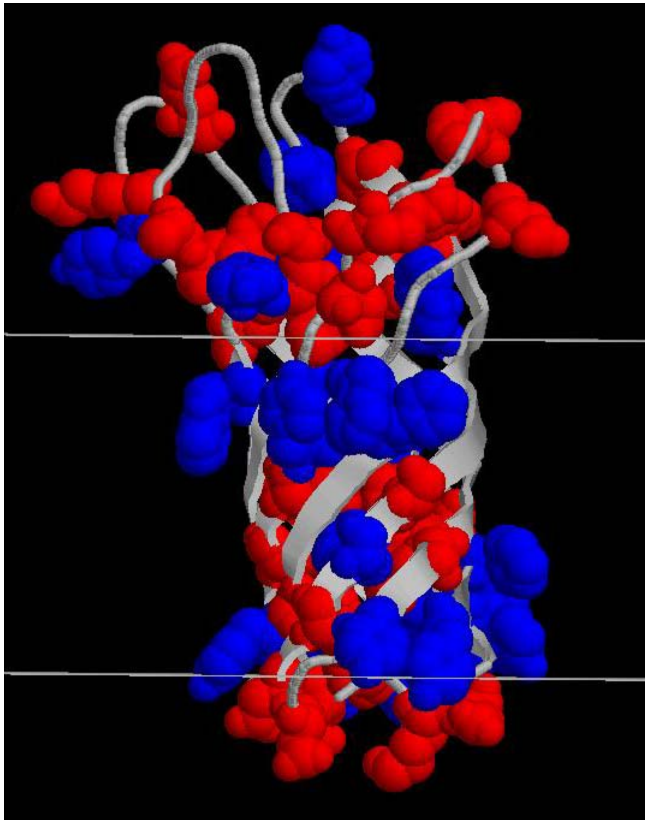
- Hypergeometric distribution for *i* of type *X* in region *r* :

$$\mathbb{P}_{X|r}(i) = \frac{\binom{n_x}{i}\binom{n-n_x}{n_r-i}}{\binom{n}{n_r}} \quad \text{and} \quad \mathbb{E}[f'(X|r)] = n_r \cdot n_x / n$$

- *p*-value for observing $f(X|r)$:

$$p = 2 \cdot \sum_{i=0}^{f(X|r)} \mathbb{P}_{X|r}(i)$$

High Propensity Single-Body Propensity



Internal headgroups and core are similar:
Glu (1.60), Arg (1.59), Gln (1.56)

Asn (1.56), **Lys (1.54), Arg (1.54)**

Trp (3.63), Tyr (2.91), Phe (1.89)

Val (2.65), Leu (2.49), Ile (2.29)

Phe (2.70), Trp (2.54), Tyr (2.46)

Ext.

Pro (2.79), Asp (1.83), Phe (1.41)

- Aromatic residues form girdles
- Polar: internal facing
- Pro in short loops

Arg and Lys enriched in extracellular cap regions.
— positive outside rule.

Positive-outside Rule

- Opposite to helical membrane proteins.
  - Positive inside rule: the most powerful topology determinant (von Heijne).

- Possible reason: asymmetric distribution of lipids in outer membrane
  - Inner leaflet facing periplasm: PE (phosphatidylethanolamine)
  - Outer leaflet facing extracellular environment: LPS (lipopolysaccharides, wit negatively charged groups).

- Hypothesis: Determinant for membrane insertion?
  - Different affinity of two sides of the barrel for the two leaflets of outermembrane.
  - Interaction with LPS before co-inserting into membrane.

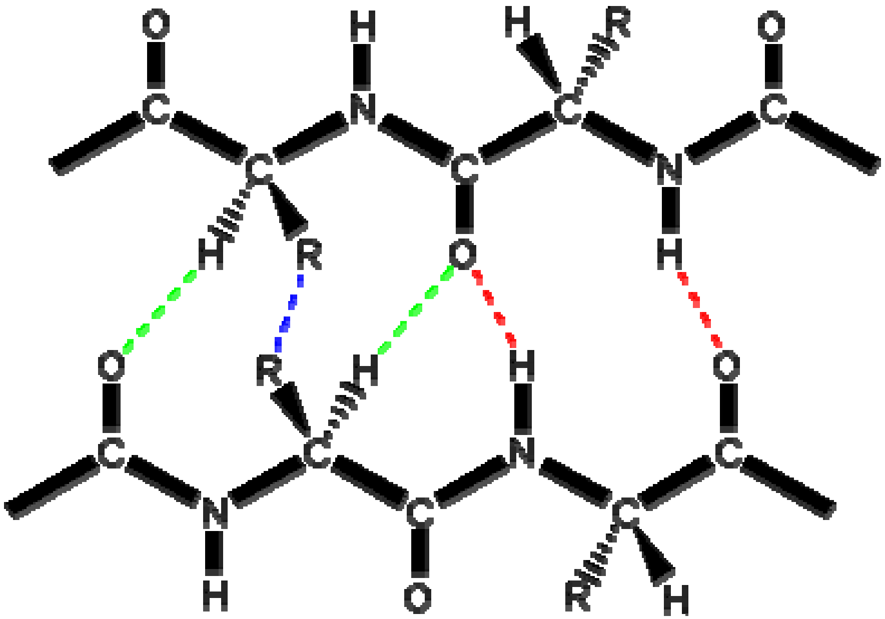
Strand-strand interactions

(Jackups, Jr and JL, *J Mol Biol*, 2005)

Interaction Pattern of Antiparallel β-Sheets

Adjacent strands:

1. Strong H-bonds immediately across
2. Non-H-bond interactions
3. Weak C-O H-bonds across and one residue displaced on the strand



(Soluble and membrane proteins. Ho and Curmi, 1999, JMB)

Membrane Strand Interface Pair propensity (MSIP)

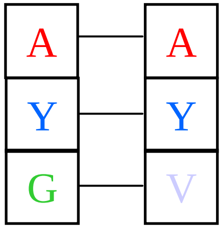
- Interstrand contacts between *X* and *Y* residues:

  MSIP :

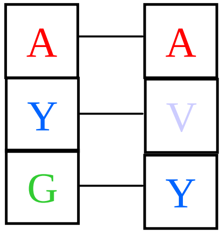
$$P(X,Y) = \frac{f(X,Y)}{\mathbb{E}[f'(X,Y)]}$$

- Null model:
  - Two adjacent strands are permuted exhaustively and independently.
  - Each permutation equally likely.

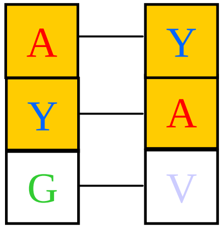
Example: AY contacts in 3-residue strand pair



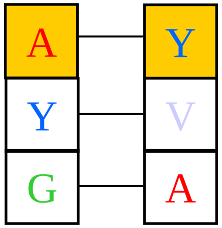
0 contacts



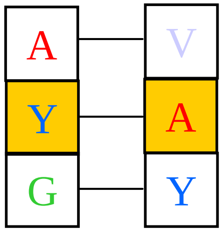
0 contacts



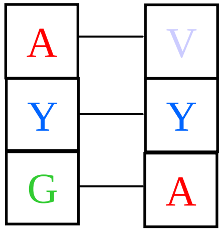
2 contacts



1 contact



1 contact



0 contacts

Expected number of AY contacts for this strand pair = 4/6 = 0.66.

Contacts between residues of same type

- Probability for *i* number of X-X contacts in pairs of length *l*:

$$\mathbb{P}_{X,X}(i) = \frac{\binom{x_1}{i}\binom{l-x_1}{x_2-i}}{\binom{l}{x_2}},$$

- Expected number of X-X contacts for all strand pairs:

$$\mathbb{E}_{\mathsf{all}}[f'(X,X)] = \sum_{sp\in\mathcal{SP}} \mathbb{E}_{sp}[f'(X,X)] = \sum_{sp\in\mathcal{SP}} \frac{x_1(sp)\cdot x_2(sp)}{l(sp)},$$

- *p*-value can calculated analytically.

$\mathbb{P}_{X,X}$(i) is hypergeometric.

Proof.

- Let there be x_1 X residues in strand 1 and x_2 X residues in strand 2 in a strand pair of *l* residues.

- Fix strand 2 and permute strand 1. Select x_2 residues from strand 1 to interact with the x_2 X residues in strand 2. *i* of these must be selected from the x_1 X residues in strand 1, and x_2-*i* of these must be selected from the *l*-x_1 non-X residues in strand 1. This is:

$$\binom{x_1}{i}\binom{l-x_1}{x_2-i}$$

- Dividing by the number of ways to permute strand 1, the result is hypergeometric:

$$\mathbb{P}_{X,X}(i) = \frac{\binom{x_1}{i}\binom{l-x_1}{x_2-i}}{\binom{l}{x_2}},$$

Contacts between residues of different type

- Expected frequency of X-Y contacts in all strand pairs:

$$\mathbb{E}[f'(X,Y)] = \sum_{sp \in \mathcal{SP}} \{\mathbb{E}[f'_{sp}(X,Y)] + \mathbb{E}[f'_{sp}(Y,X)]\}$$

$$= \sum_{sp \in \mathcal{SP}} \{\frac{x_1(sp) \cdot y_2(sp)}{l(sp)} + \frac{y_1(sp) \cdot x_2(sp)}{l(sp)}\},$$

- *p*-value is more difficulty, as f'$_{sp}$(X,Y) and f'$_{sp}$(Y,X) are dependent.

Calculating p-value of X-Y contacts: Generalized Hypergeometric Distribution

- Trinomial function (a,b,c)! = (a+b+c)!/a!b!c!.
- Define: $T(l, x_1, y_1) \equiv (x_1, y_1, l-x_1-y_1)!$
- The random probability of *h* X-X contacts, *i* X-Ycontacts, *j* Y-X contacts, and *k* Y-Y contacts is:

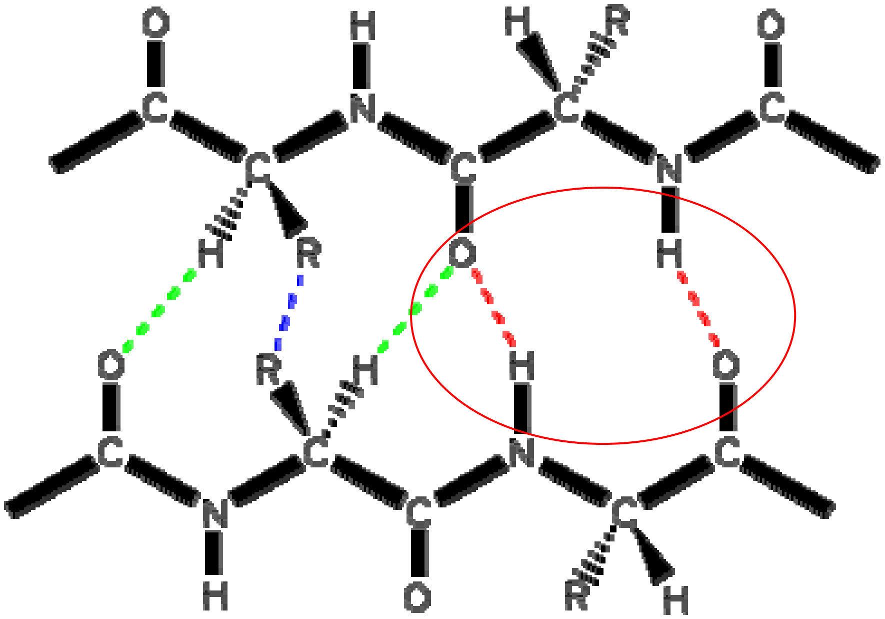
$$\mathbb{P}(h,i,j,k) = \frac{T(x_1,h,i)\cdot T(y_1,j,k)\cdot T(l-x_1-y_1, x_2-h-j, y_2-i-k)}{T(l,x_2,y_2)}.$$

- Marginal probability of a total of i+j = m X-Y contacts for computing p-value:

$$\mathbb{P}_{X,Y}(m) = \sum_{h=0}^{x_1} \sum_{i=0}^{x_1-h} \sum_{k=0}^{\substack{y_1-\\(m-i)}} \mathbb{P}(h,i,m-i,k),$$

Significant Interaction Motif and Antimotifs for Strong H-bonds

High		Low	
Pair	Odds	Pair	Odds
GY	1.56	YY	0.28
ND	2.76	WY	0.21
GF	1.80	AW	0.22
IY	1.79	PV	0.00
KS	1.95	VV	0.65
AA	1.60		
LW	1.92		
LY	1.37		
RP	4.00		
HK	3.33		
ET	1.55		
NN	1.94		
G-FWY			1.52
ILV-FWY			1.31

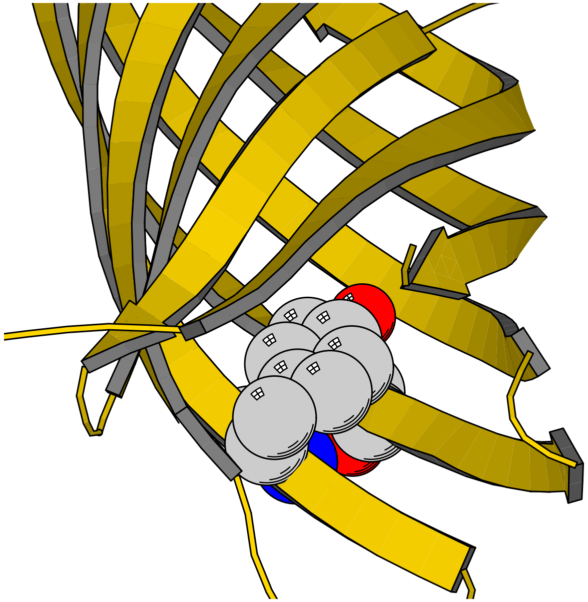


🟥 *p*-value < 0 .05

🟦 *p*-value < 0.10

Example: Aromatic Rescue

- Internal Y-G backbone H-bond interactions: 1.56, significant *p*-value (8.0×10^{-4}).
  - — Normally internal Y disfavored.
  - — 60% takes unusual (60, 90) rotamers vs 6% in soluble beta-strands



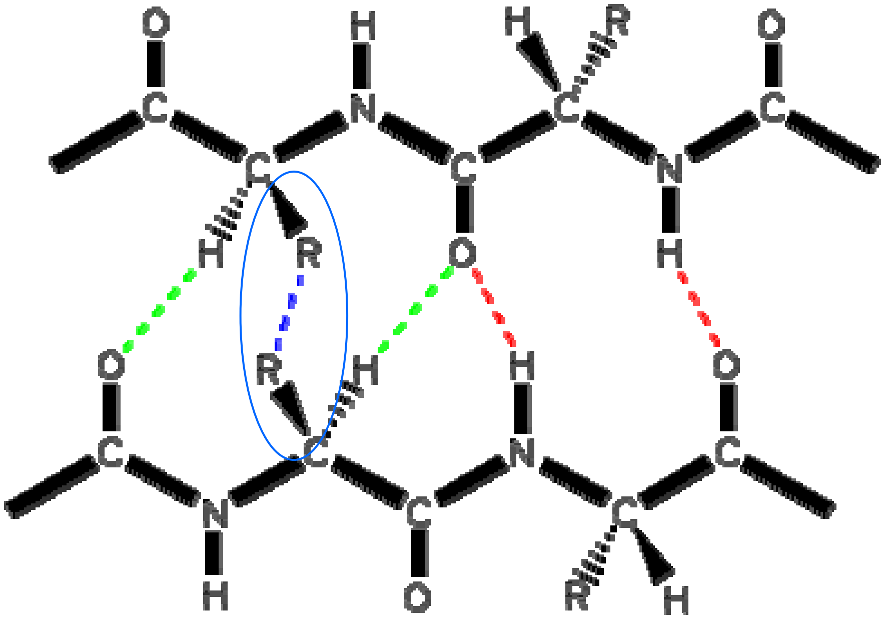
- Y covers G: mitigates instability of G causes in β-sheets.
  - — Prevents unfavorable exposure of backbone around G and aromatic ring to solvent
  - — For membrane protein: accomodating curvature
  - — Important for folding.
    - eg. Three cooperative internal G-Y for folding

(Merkel and Reagan, 1999).

Other aromatic rescues:
G-F is significant, and has the same rotamer preference..

Significant Interaction Patterns for Non-H-bond Interactions

High		Low	
Pair	Odds	Pair	Odds
WY	2.71	GK	0.32
GI	1.77	QV	0.00
RE	1.87	GY	0.62
GV	1.60	NL	0.33
QG	1.57	QI	0.00
LL	1.44	AT	0.59
AA	1.54	IF	0.42
LP	2.07		
AV	1.39		
FY	1.50		
QK	2.09		
NS	1.58		
FWY-FWY			1.48
G-ILV			1.48



🟥 p-value < 0.05

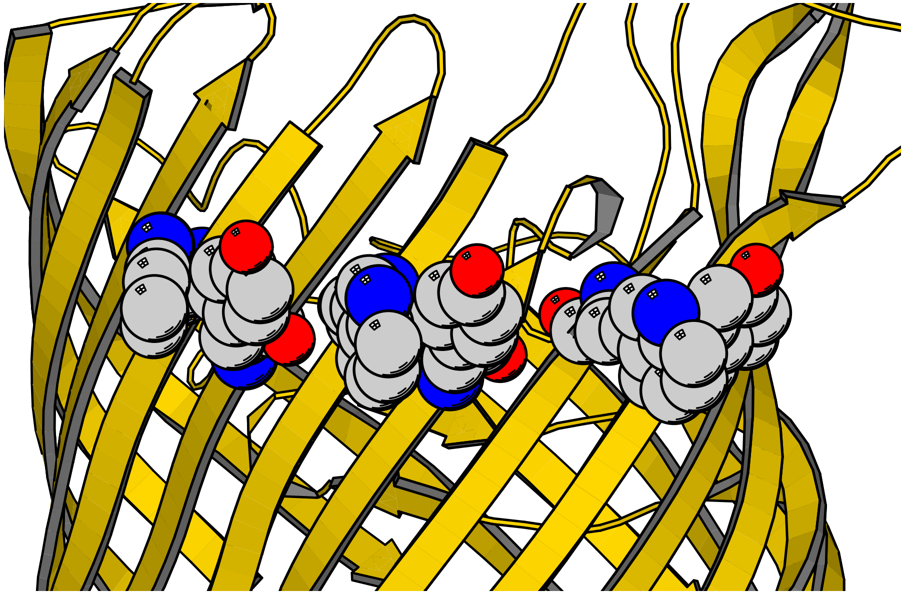
🟦 p-value < 0.10

Example: Trp-Tyr Side Chain Interactions

- W-Y side chain interactions: significant *p*-value (4 ×10 $^{-7}$).
- Maybe responsible for rotamer bias of W and Y.

W and Y have unique rotamer preferences in TM β sheets

(Chamberlain and Bowie, 2004).



Three WY side chain interactions in maltoporin.

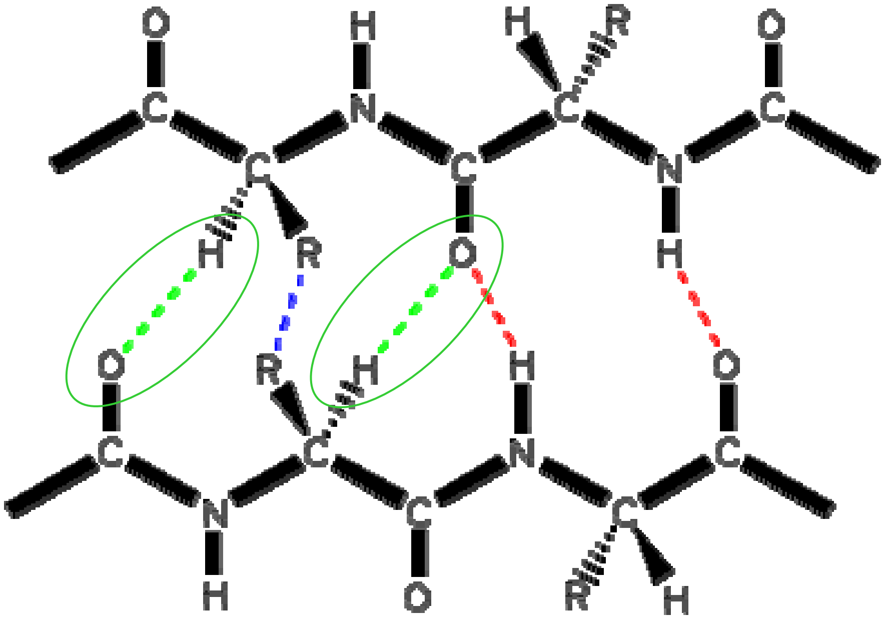
W prefers a +60° rotation.
Y prefers a +180° rotation.

Significant Interaction Patterns for Weak H-bonds

High		Low	
Pair	Odds	Pair	Odds
DV	1.98	FV	0.07
NI	2.24	VV	0.13
GL	1.39	ES	0.00
GP	2.37	RD	0.00
DP	4.10	IL	0.44
NV	1.75	LV	0.60
LS	1.47	NG	0.50
RF	1.78	DS	0.32
IS	1.69	QG	0.54
EL	1.58	ND	0.00
AK	1.60		
EF	1.83		
ILV-polar			1.39
FWY-polar			1.39

🟥 p-value < .01

🟦 p-value < .03



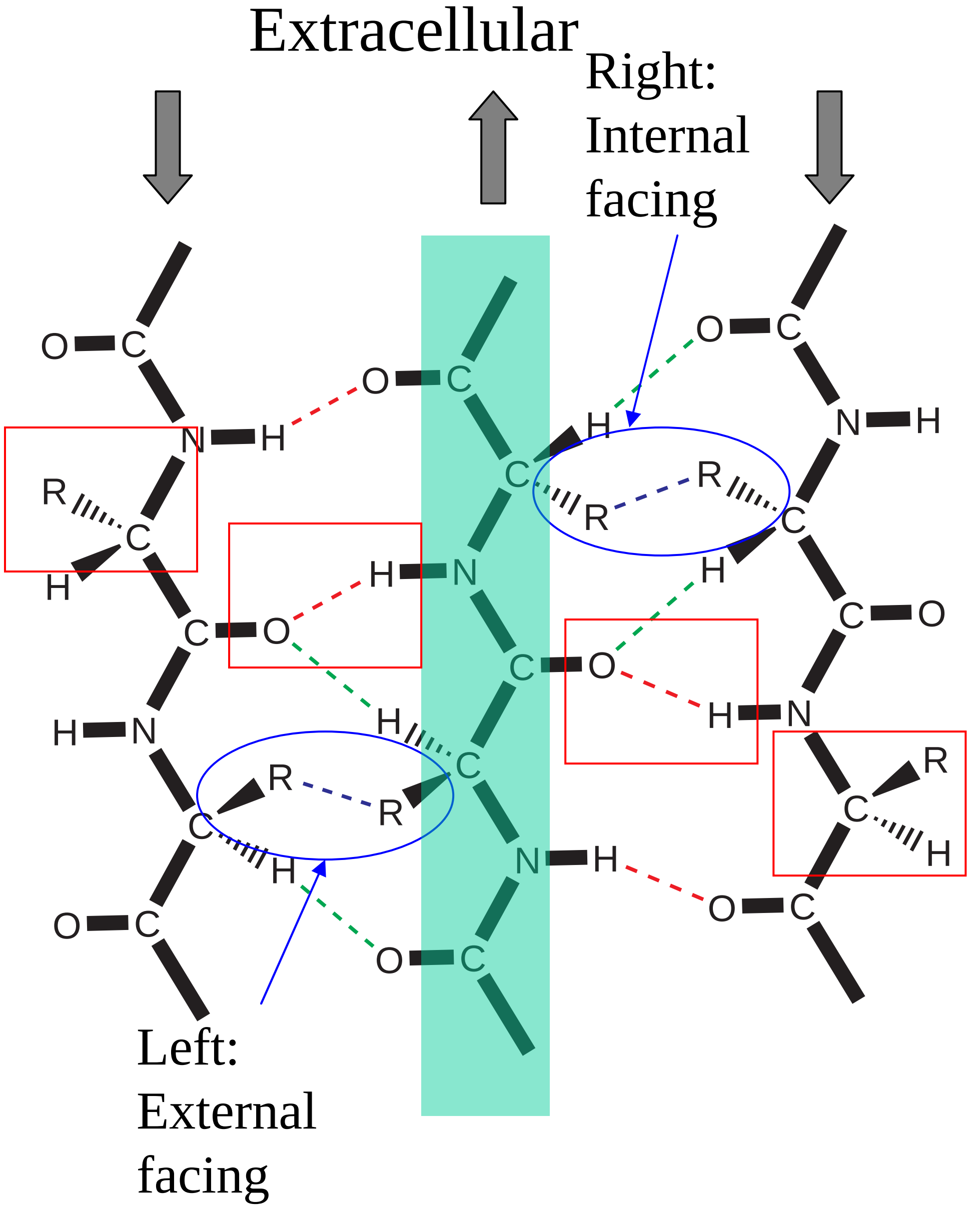
- Two residues involved in weak H-bonds face opposite directions.
- Pairwise propensities may be confounded by single-body preferences.

Patterns due to Chirality

(Up-Strand: N to C towards extracellular side)

- *L*-amino acids: side chains
  - Sidechain to the “**right**” of up-strands always internally facing.
  - Sidechain to the “**left**” of up-strands always externally facing.

- *L*-amino acids: strong H-bond
  - Reverse patterns.



Extracellular

Right: Internal facing

Left: External facing

Sequence Motifs

(R Jackups, Jr, S Cheng, and JL, manuscript)

- Helical proteins: *Senes*, *Gerstein*, *and Engelman*, *2000*, *JMB*.
  - GxxxG in GpA
  - Many other motifs: GA4, etc.
  - Important for TM helices assembly and for dimerization.

- **Problem:** Given a database of helices/strands, are there motifs of two residues in the form of i (k-1) j ?

  GxxxG, GxxxA

  - Odds ratio of **frequencies of observed** against **expected frequency** by random permutation of residues.
  - Significance level.

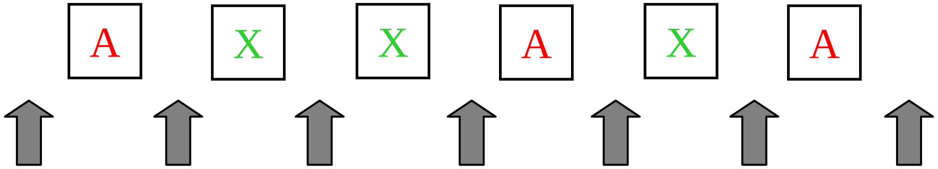
- Data: 15,946 predicted transmembrane β-strands (Bigelow et al.).

Expected probability *p*(*i*,*j*,*k*) of *i* (*k*-1) *j* motifs

p(i,j,k) is hypergeometric if *k* = 1

Proof: Let there be *m* A residues and *n* Y residues in a strand of *l* residues. Determine *P(x|l,m,n)*, the probability of *x* instances of AY1 motifs

First, place the *l*-*n* non-Y residues arbitrarily:

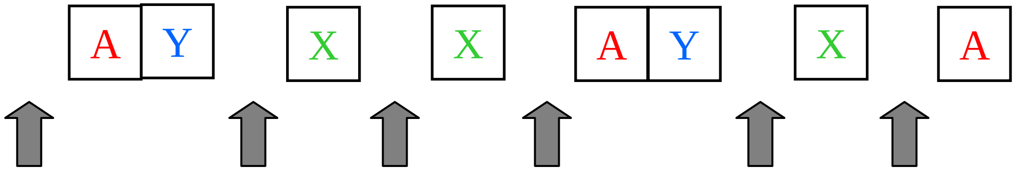


$l = 10$
$m = 3$
$n = 4$

- Total number of ways that the *n* Y residues can be placed in the *l*-*n*+1 possible slots above, *with replacement*, is:

$$\binom{(l-n+1)+n-1}{n} = \binom{l}{n}$$

- Next, select *x* of the *m* A residues after which a Y will be placed:



l = 10
m = 3
n = 4
x = 2

- The number of ways this can be done, multiplied by the number of ways the remaining *n*-*x* Y residues can be placed, *with replacement*, in the *l*-*n*+1-*m*+*x* slots that do not immediately follow an A, is:

$$\binom{m}{x}\binom{(l-n+1-m+x)+(n-x)-1}{n-x}=\binom{m}{x}\binom{l-m}{n-x}$$

- This gives a hypergeometric:

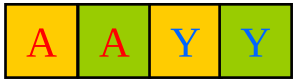
$$p(x\,|\,l,m,n)=\frac{\binom{m}{x}\binom{l-m}{n-x}}{\binom{l}{n}}$$

How to Identify Sequence Motifs?

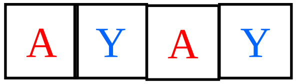
- Analytical formula when *k* =1.

- Enumerate all possible permutations of a strand (*Senes et al*).
- Examine how often a pattern occurs.

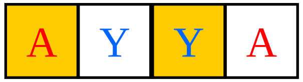
Example: AY2 motif in 4-residue strand


A A Y Y

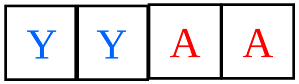
2 occurrences


A Y A Y

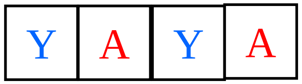
0 occurrences


A Y Y A

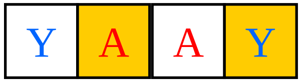
1 occurrence


Y Y A A

0 occurrences


Y A Y A

0 occurrences


Y A A Y

1 occurrences

Expected number of AY2 occurrences for this strand:
4/6 = 0.66

How to Identify Sequence Motifs using a Database?

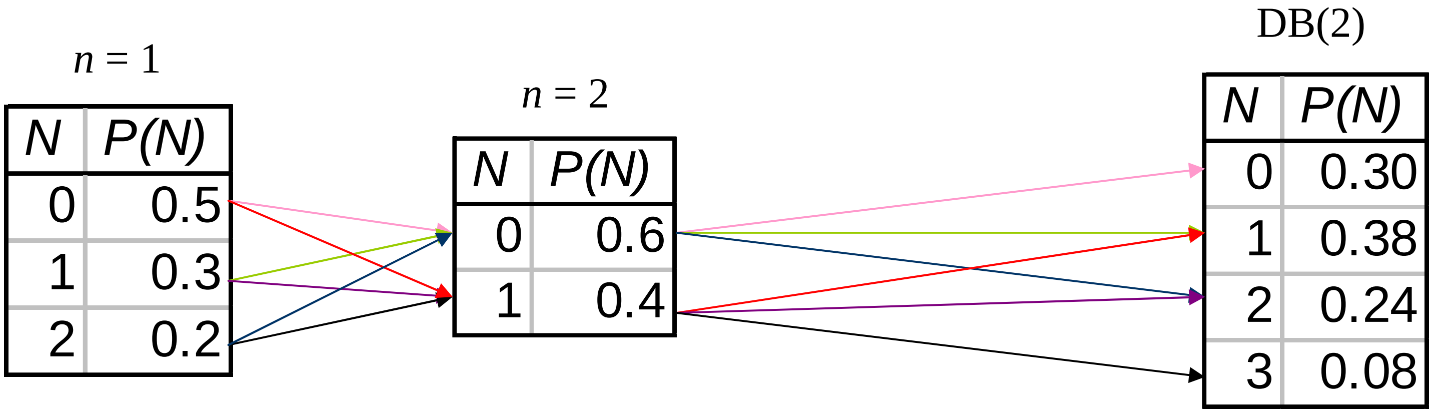
- Expected probability $p(i,j,k)$ of i (k) j motifs:
  - — Count all i (k) j motifs in the exhaustively permuted dataset.
- Dynamic programming algorithm designed by Senes *et al*.

$$p_{DB(n)}(N_{i,j,k}) = \sum_{l=0}^{N_{i,j,k}} p_{DB(n-1)}(l_{i,j,k}) p_n(N_{i,j,k} - l)$$

$p_{DB(n)}(N_{i,j,k})$: probability of N of i (k) j motifs in a database of n strands

$p_n(N_{i,j,k})$: probability of N of i (k) j motifs in n^{th} strand of database

Example: Combining three distributions into one



n = 1

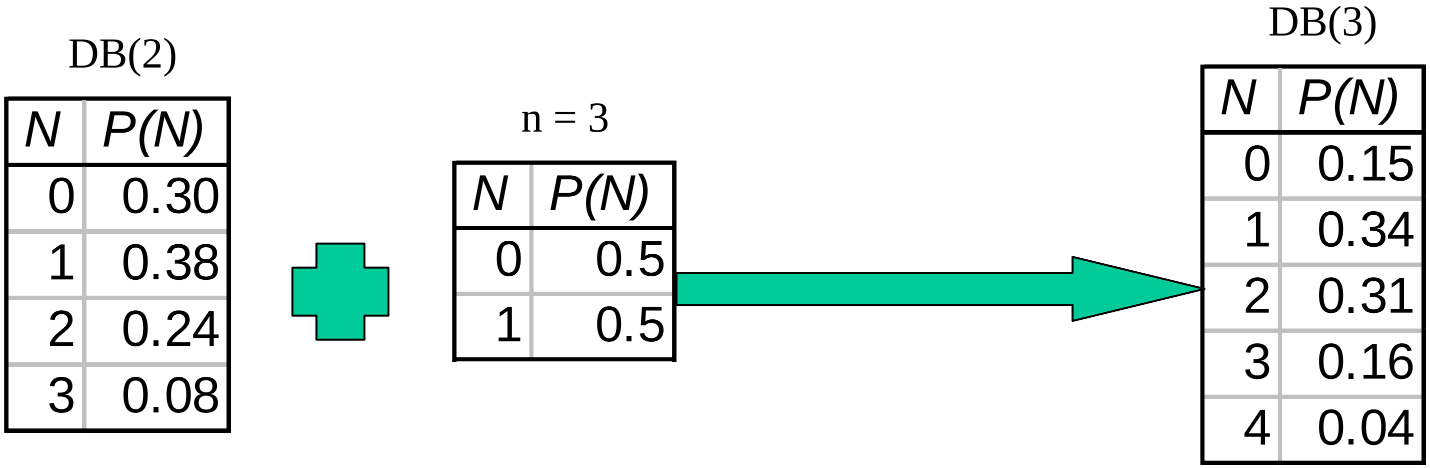
N	*P(N)*
0	0.5
1	0.3
2	0.2

n = 2

N	*P(N)*
0	0.6
1	0.4

DB(2)

N	*P(N)*
0	0.30
1	0.38
2	0.24
3	0.08



DB(2)

N	*P(N)*
0	0.30
1	0.38
2	0.24
3	0.08

n = 3

N	*P(N)*
0	0.5
1	0.5

DB(3)

N	*P(N)*
0	0.15
1	0.34
2	0.31
3	0.16
4	0.04

All Significant Sequence Motifs and Antimotifs

Motif	Odds	p-value
LY6	2.74	7E-266
LY4	2.31	2E-260
GY3	2.19	1E-242
WY8	7.24	8E-195
AY2	1.93	6E-177
LL2	1.61	4E-159
WY6	4.38	5E-157
FY6	2.95	5E-153
LL1	0.53	6E-144
VY6	2.54	4E-134
VY4	2.09	4E-117
LA2	1.56	2E-109
LR2	0.39	6E-102
RD2	2.40	2E-100
FL1	0.40	3E-98

🟥 high propensity

🟦 low propensity

1. Preliminary: sequence database needs clean up.
2. Correlated with regional preference of residues.

eg. Exterior facing residues when *k* is even.

Significant Sequence Motifs with *k* = 2

Motif	Odds	p-value
AY2	1.93	6E-177
LL2	1.61	4E-159
LA2	1.56	2E-109
LR2	0.39	6E-102
RD2	2.40	2E-100
LY2	1.63	1E-97
RL2	0.42	2E-92
LD2	0.41	1E-80
LN2	0.45	2E-79
LE2	0.36	2E-76
VY2	1.70	2E-74
VL2	1.51	4E-61
SL2	0.59	4E-60
VA2	1.51	2E-56
LV2	1.48	1E-53

Physical basis: often nearest neighbor side chain interactions.

🟥 high propensity

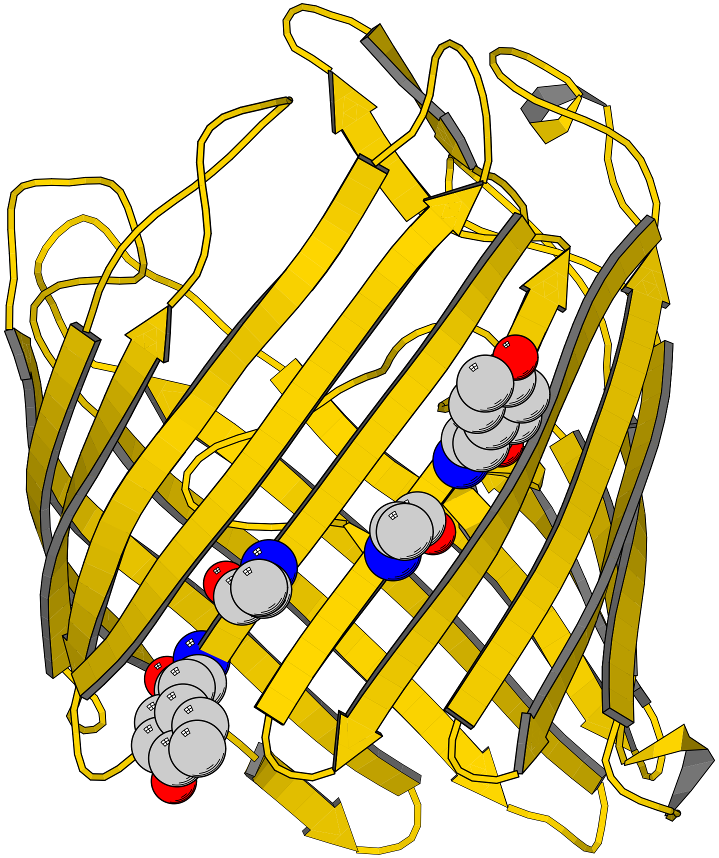
🟦 low propensity

Example: Dichotomy of Ala-Tyr-2 Motifs

- AY2 motif:
  - High propensity (1.93), and very significant *p*-value (6× $^{-177}$),
- YA2 is an anti-motif:
  - Low propensity (0.69, *p*-value = 3 × $^{-27}$).
- Similar results with only 18 protein structures only.
  - AY2: 1.90 (*p*-value = 4 × $^{-4}$)
  - YA2: 0.39 (*p*-value = 6 × $^{-3}$)
- Y: unique rotamer preference in transmembrane β-sheets.

(Chamberlain and Bowie, 2004)

- Others: VY2-YV2, GY3-YG3, LY4-LY4



Two AY2 motifs in OmpF
Y prefers a +180° rotation.

Motifs in loop region

- By enumeration and dynamics programming.
- Exact *p*-value impossible, but can be approximated.

- Experimental data: Periplasmic chaperon preferrably binds to Aromatic-x-Aromatic, Aromatic-x-Pro motif.
  - This motif is not found in TM strands.

(Bitto and McKay, 2003, JBC)

- The only favorable motif in loops: YF2
  - May be chaperon binding site.
- Other site: WP2 marginally favorable.

Biological Significance of Discovered Patterns and Motifs

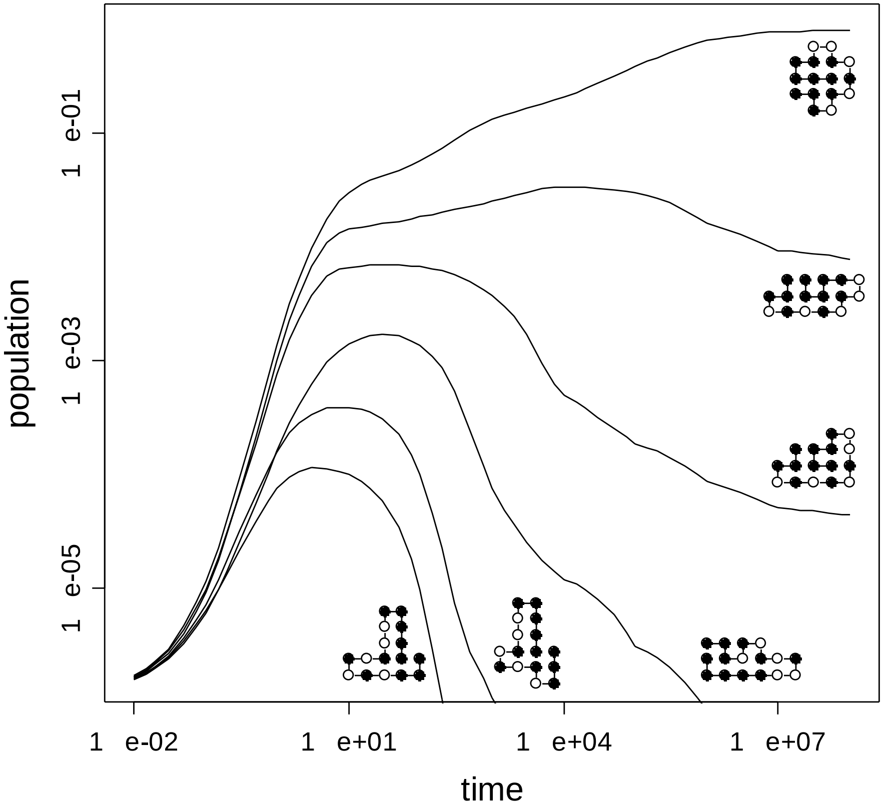
- Folding and assembly
  - Depend on the thermodynamics of lipid, proteins, and their interactions.
- Sorting and targeting process for biological localization
  - Complex biomolecular machinery, eg. translocon, chaperon.

Suggested Experiments

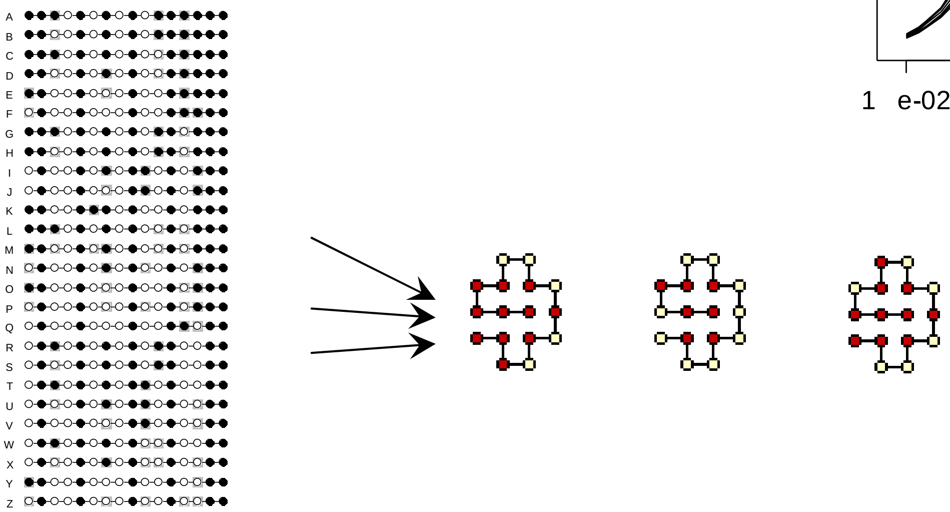
- Outside positive rule:
  - Introducing Arg and Lys in periplamic cap region and test for folding.
    - Discriminating mechanism: In vivo sorting or folding
  - Reconstitute lipid bilayer of different compositions
    - Whether origin of this rule is due to assymetric lipid distribution.
- Aromatic rescue and other motifs:
  - Gly-Tyr may be anchoring site for folding.
  - Remove or add Gly-Tyr for folding rate studies.

Further Studies

- Folding dynamics and mechanisms.
  - Models of physical interactions.
  - Effective potential functions.



- Structure prediction.



Master equation, matrix exponential and Krylov subspace method.

(Sëma Kachalo, Hsiao-Mei Lu, and JL, *Phys Rev Lett*, 2006, 96: 058105.1-4)

Structure Prediction: strand-pairing

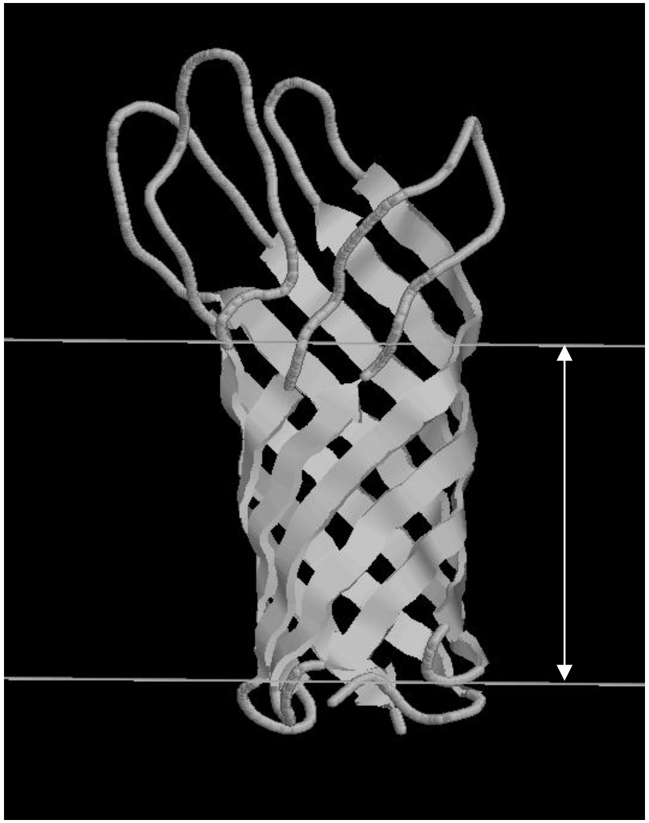
β-barrel membrane proteins

(with Ronald Jackups)

Architecture of β-Barrel Membrane Proteins

- β-barrel membrane proteins are twisted and tilted into barrels
  - Right-hand twist
  - Tilt ≈ 30-60º

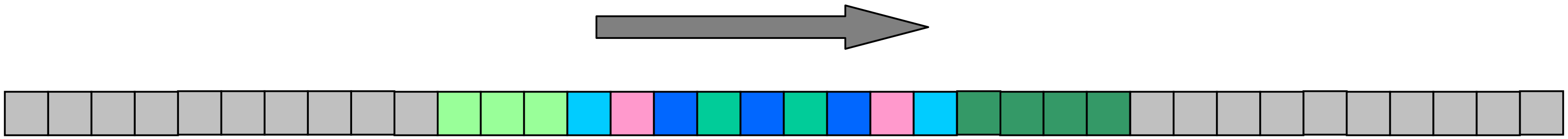
- Loops extend outside membrane
  - Short Pro-rich periplasmic loops
  - Long polar extracellular loops



Arrow: Width of outer membrane (~27 Å)

Predicting structures of β-barrel membrane proteins

- Need to predict register pattern of adjacent strands.
- Enumerate all possible registers.



Single-Body Propensities:

🟩🟩 Cap regions (outer and periplasmic)

🟦🟥 Interface regions (inside and out)

🟦🟩 Core region (inside and out)

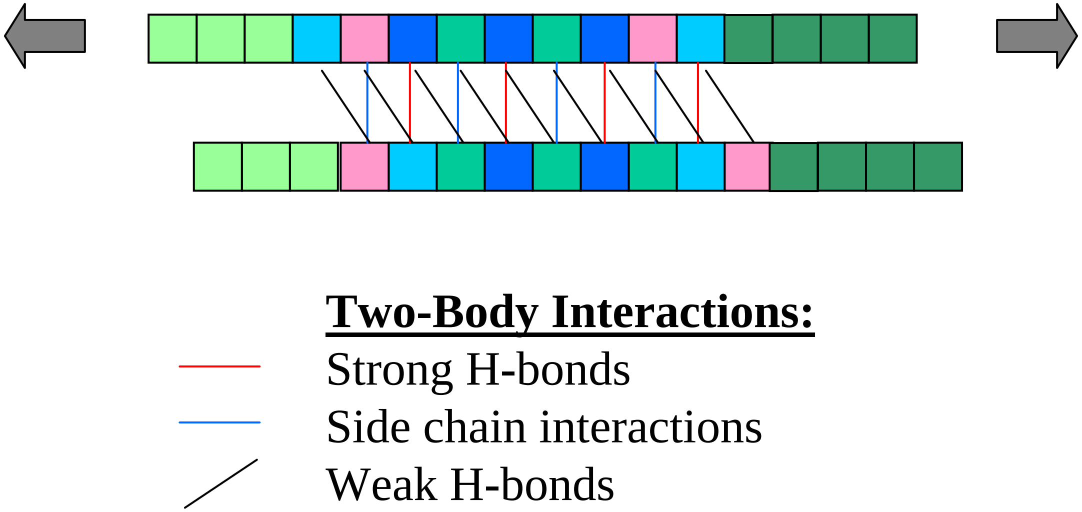
1. Thread sequence through to find highest scoring strand.

  - Strand: windows of 16 residues,
  - Use single-body preference.

– Other methods: HMM (Bigelow et al).

2. Find the best register between two strands:

— Evaluation of energy score.



Two-Body Interactions:

Strong H-bonds

Side chain interactions

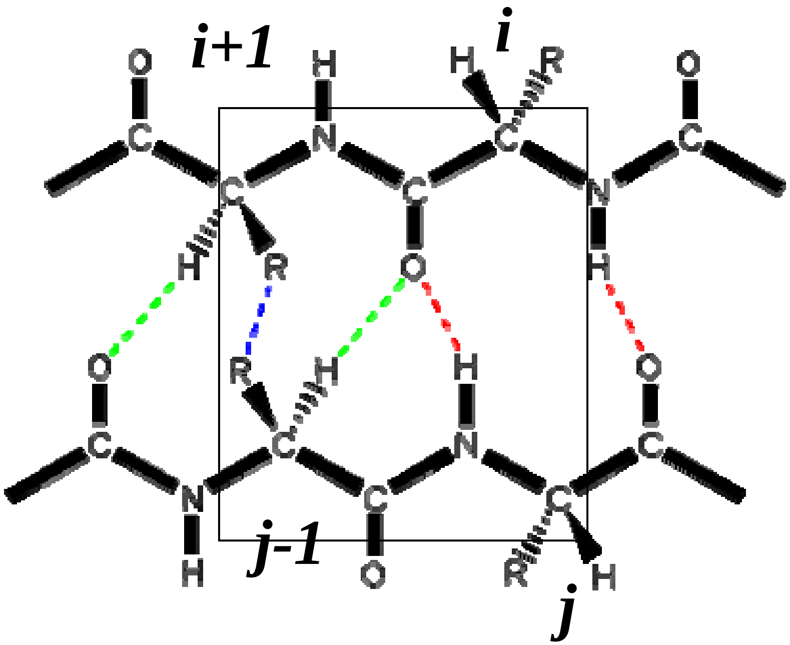
Weak H-bonds

— Chirality reduces search space by ½.

Energy Score

- Energy from two interacting pairs:

$$\begin{aligned} E_I(i,i+1;\, j\text{-}1,j) = \alpha_1\, [&E_1(a_i) + E_1(a_j) + E_1(a_{i+1}) + E_1(a_{j\text{-}1})] + \\ &\alpha_{NO}\, E_{NO}(a_i,\, a_j) + \alpha_{sc}\, E_{sc}(a_{i+1},\, a_{j\text{-}1}) + \\ &\alpha_{C\alpha}\, E_{C\alpha}\, (a_i,\, a_{j\text{-}1}) \end{aligned}$$



$$E = -\ln p$$

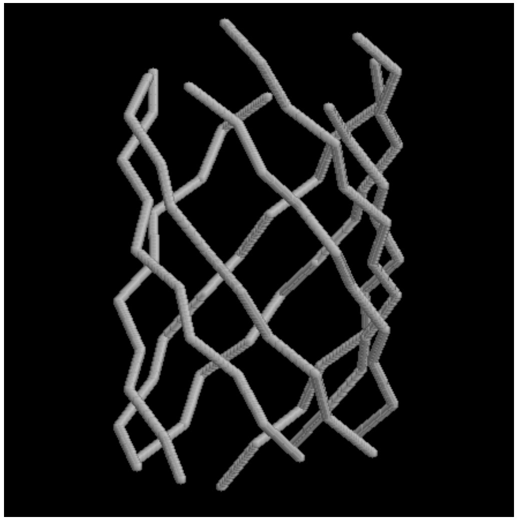
Prediction Results

- Leave-one-out test of 18 proteins.
  - Extracting pattern and *p*-values from 17 structures.
  - Predicting the remaining 1.
  - Take turns.
- Strand starts unknown:
  - Accuracy: 45%
  - Random: 5%
- Strand starts known:
  - Accuracy: 64% vs 31% by Thornton.
  - Random: ~30%

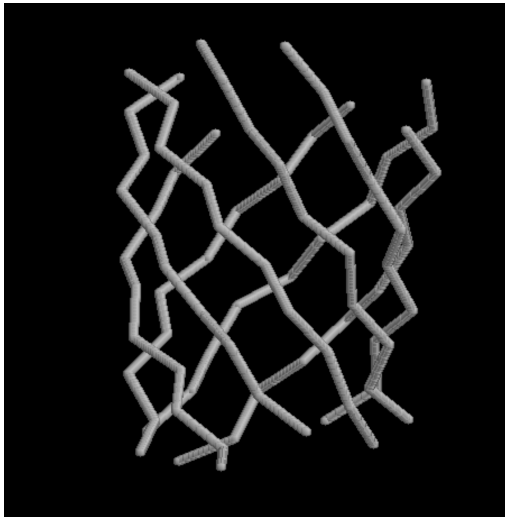
		Strand	**Strand**
PDB	**#Strands**	**Unkown**	**Kown**
1A0S	18	5	9
1BXW	8	3	4
1E54	16	7	11
1FEP	22	10	15
1I78	10	3	5
1K24	10	3	6
1KMO	22	16	18
1NQE	22	7	16
1P4T	8	4	8
1PRN	16	8	10
1QD6	12	4	5
1QJ8	8	5	3
1UYN	12	9	10
2FCP	22	11	14
2MPR	18	7	11
2OMF	16	4	8
2POR	16	9	12
Total	256	115	165

Example: *Ab initio* Structure Prediction of Neisserial Surface Protein A (NspA)

- Bacteria for meningitis and septicaemia , homolog of Opa proteins
  - Adhesion to host cells (?)
- All 8 strand pairs are predicted correctly.
- RMSD = 2.5 Å for 80 transmembrane α-carbons.



Predicted Structure



Actual Structure, NspA

What are driving forces for β-barrel proteins?

Optimized weight coefficients to separate native and mismatched strand pairs.

$$E\,(i,i+1;\, j-1,j) = \alpha_{NO}\, E_{NO}(a_i, a_j) + \alpha_{sc}\, E_{sc}(a_{i+1}, a_{j-1}) + \alpha C_\alpha\, EC_\alpha\, (a_i, a_{j-1}),$$

$$\alpha_{sc} = 1.00, \quad \alpha_{NO} = 0.88, \quad \alpha_{C\alpha} = 0.17.$$

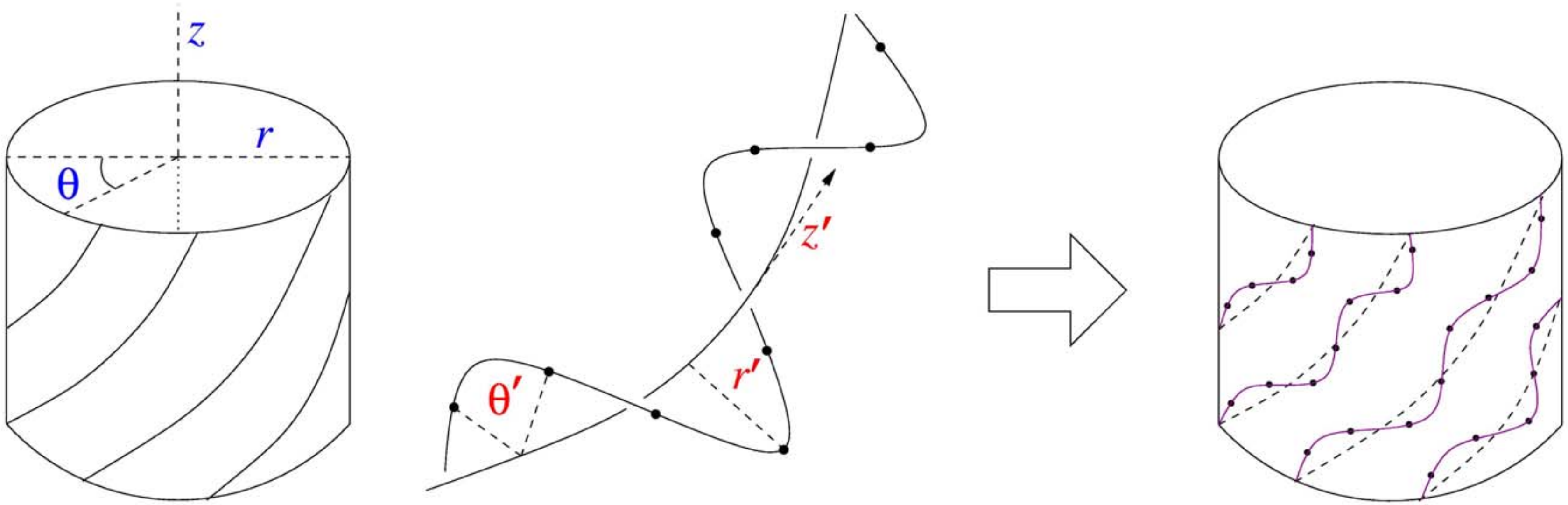
- Side chain interactions and strong H-bond are both important.
- Weak H-bond do not seem to be important.

Atomic Structure Prediction

- Parameters of strand-pairing patterns of β-barrel membrane proteins:
  - Number of strands (*N*)
  - Shearing number (*S*)
  - **Residue-specific pairing pattern**

- Task: Given only the strand-paring pattern, can we construct reliable 3-D β-barrel structures at the atomic level?
  - Main C_α trace
  - Backbone atoms (N, C=O, and C_β) contacting C_α
  - Sidechain atoms contacting C_β

Parameters of Coiled-Coils



Major Helix

r: distance from central axis of barrel

z: distance along central axis of barrel

θ: angular displacement around barrel

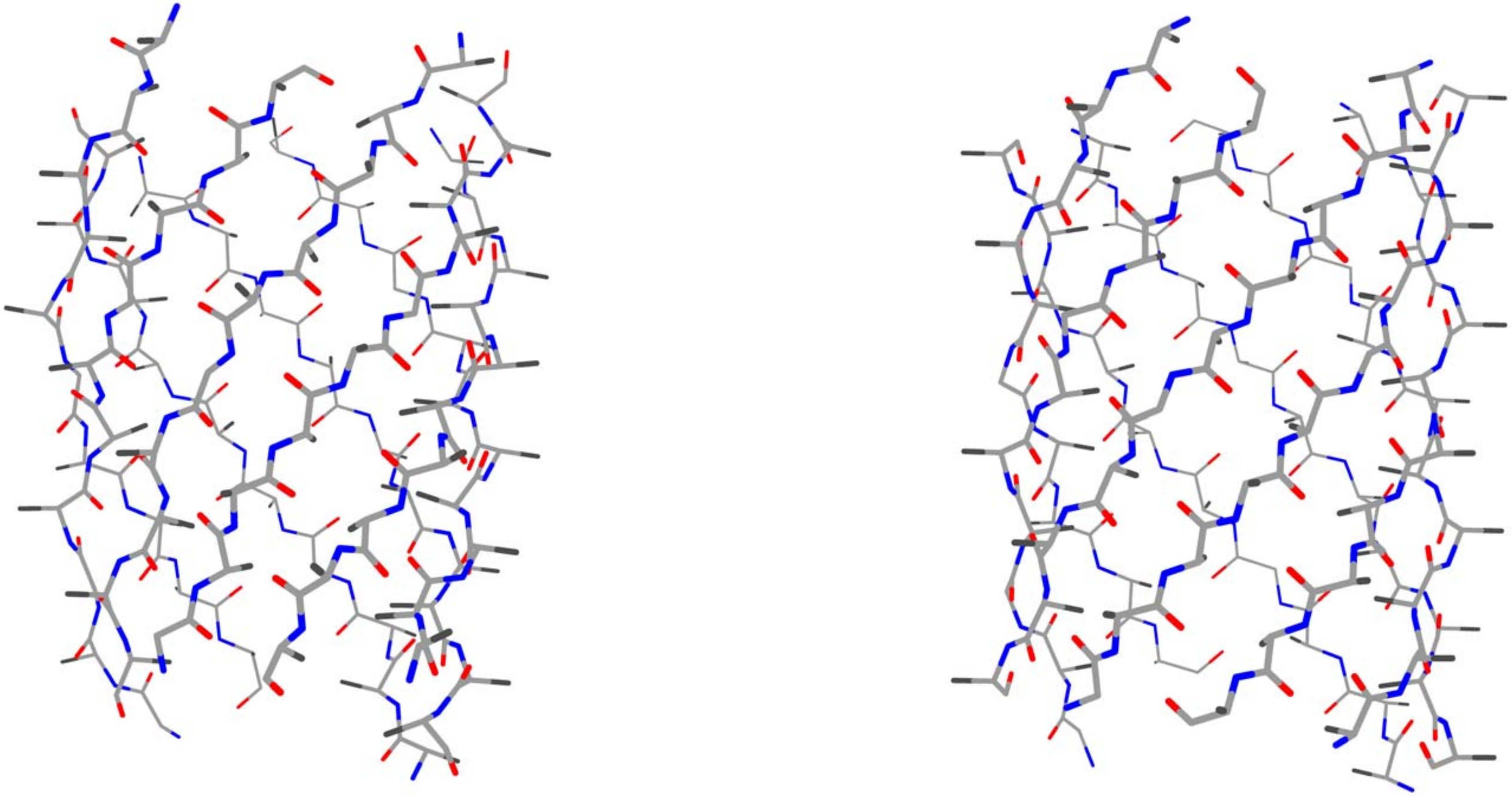
Minor Helix

r’: distance from central axis of strand

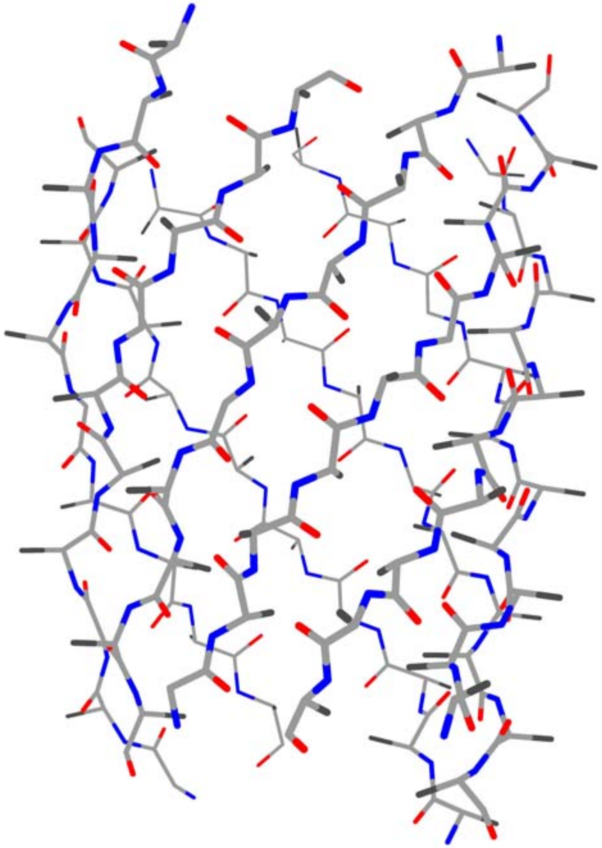
z’: distance along central axis of strand

θ’: angular displacement around strand

Example: NspA

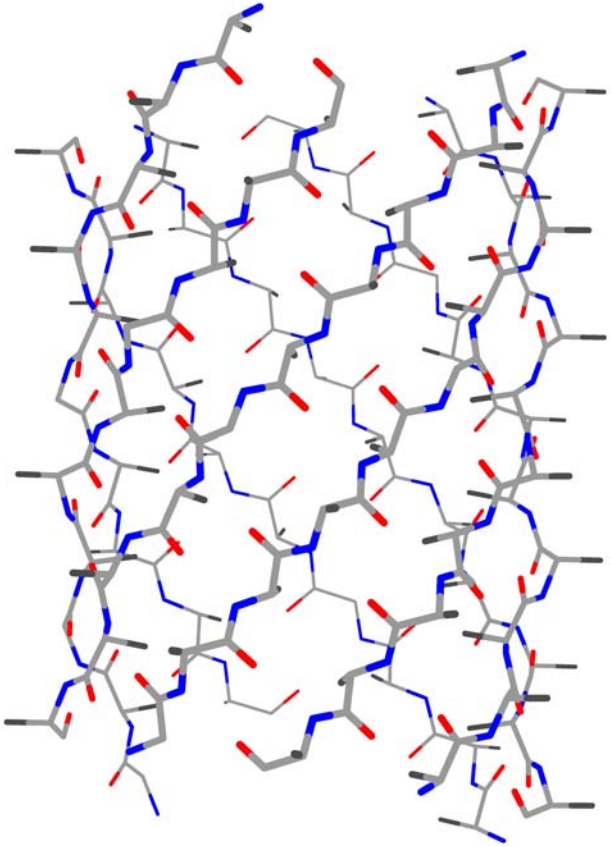


Example: NspA



True structure of NspA (PDB `1p4t`)

RMSD = 2.35
for 391 atoms



Predicted structure based on *N*, *S*, and H-bonding pattern

Results

- When *N*, *S*, and the H-bonding pattern are known, very reliable atomic ransmembrane barrel structures can be constructed
- Median RMSD = 3.34 Å and range = (2.04, 5.59) Å
  - RMSD smallest for small barrels
  - RMSD slightly higher for large metal ion transporters (which have very regular structures)
  - RMSD highest for mid-sized porins (which have irregular structures, usually due to protein-protein interfaces)

PDB	RMSD	Atoms	Res.	PDB	RMSD	Atoms	Res.	PDB	RMSD	Atoms	Res.
1bxw	2.04	385	80	**1qd6**	3.70	589	120	**1a0s**	5.59	883	180
1qj8	2.69	383	80	**2por**	3.74	779	160	**1fep**	3.00	1058	215
1p4t	2.35	391	80	**1prn**	3.34	782	160	**2fcp**	3.37	1082	220
1k24	2.70	486	100	**2omf**	3.48	778	160	**1kmo**	3.46	1078	220
1i78	2.87	486	100	**1e54**	4.03	781	160	**1nqe**	2.93	1075	220
1uyn	2.45	579	120	**2mpr**	4.20	883	180				

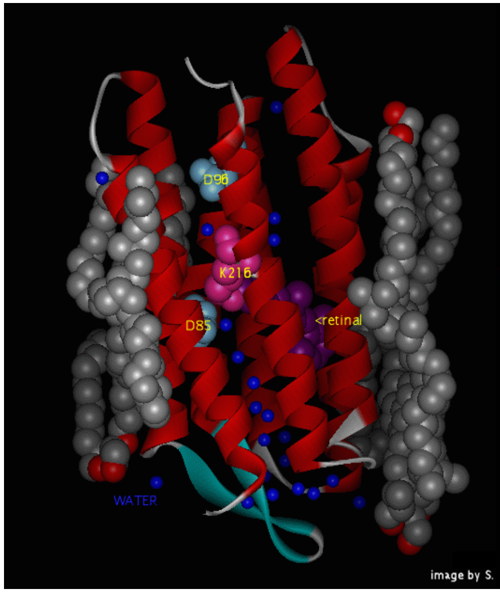
Conclusion

- Regional preference of residues can be estimated.
  - Helps to discover chemical code for membrane insertion.
- Strand pairing preferences.
  - Aromatic rescues for protein folding
  - Other stabilizing interactions.
  - Predict strand pairing
- Sequence motifs.
  - AY dichotomy: role unknown
  - Candidate chaperon binding site.
- Reliable backbone atom structures of TM β-barrels can be constructed knowing only:
  - Strand pairing

Helical Membrane Proteins

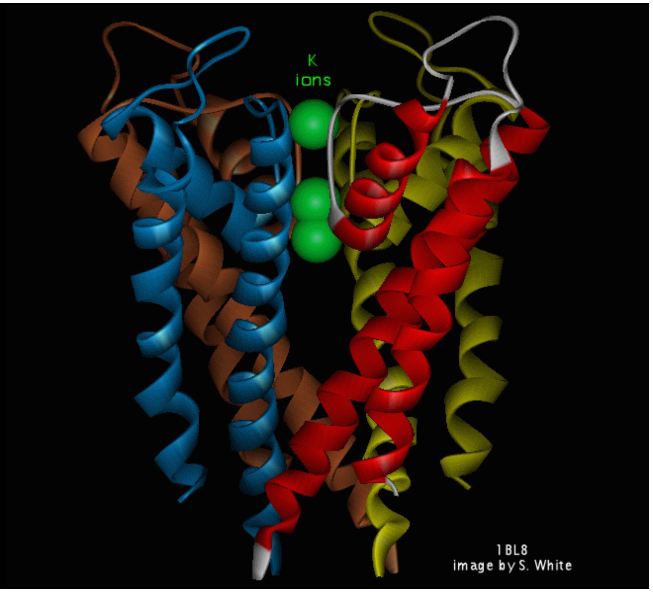
(with Larisa Adamian)

bR



rhodopsin

K-channel



(from White Lab)

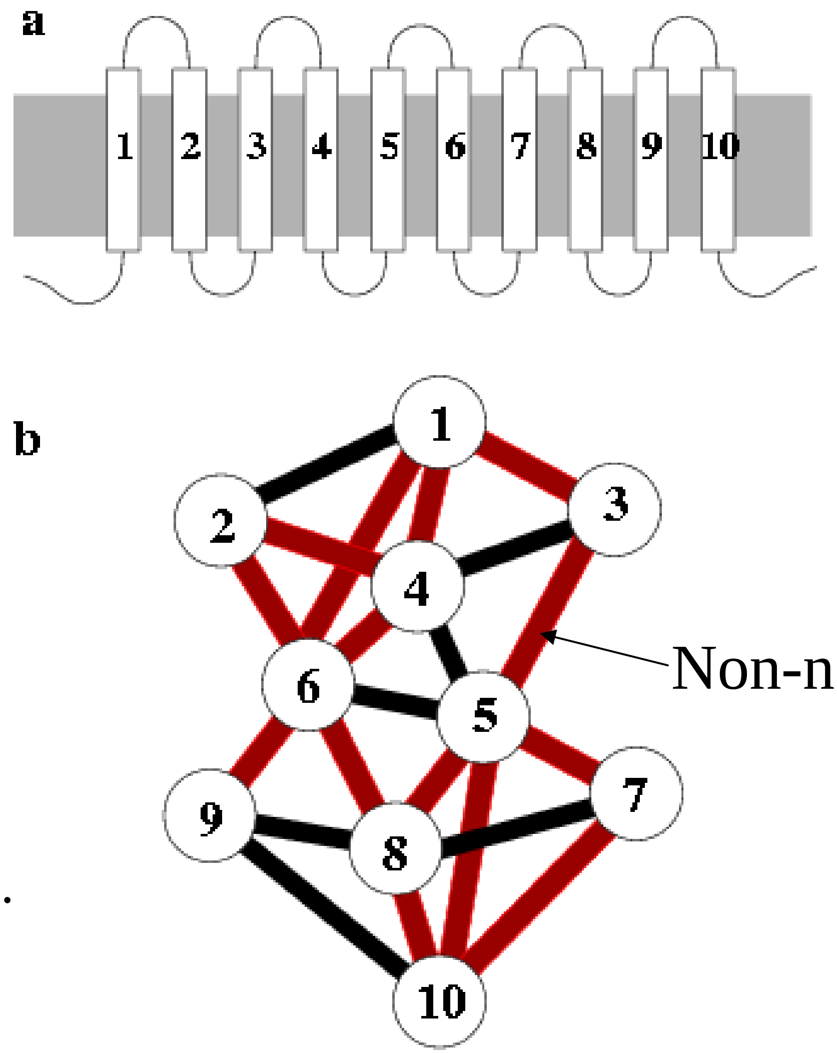
Assembly of TM Helices

- Packing : Voids and contacts
- Interhelical interactions.
  - Pairwise.
  - Higher order.
- H-bond between helices.

(Adamian and Liang, JMB, 2000; Adamian and Liang, Proteins, 2002; Adamian and Liang, JMB, 2003)

Helix-Helix Interactions

Calcium Transporting ATPase



Non-neighbors

Parallel: 10
Antiparallel: 9

(Ben-Tal & Honig, 1996)

– Often non-sequential neighbors.

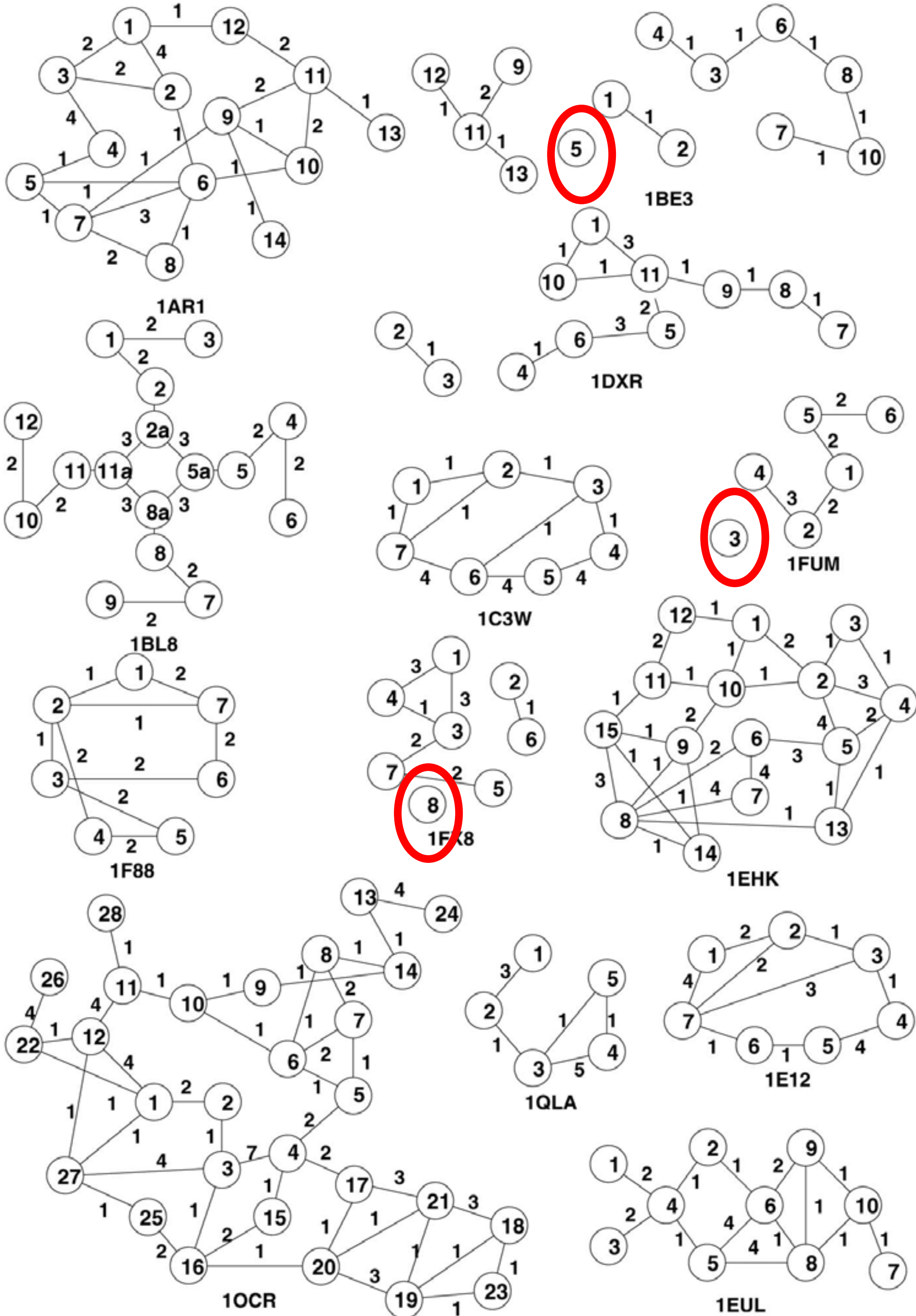
MHIP with Confidence Intervals

```
AAs  AVG    Studentized     Non-Studentized  Counts   Bias
AxA  1.29  (  0.97,  1.95)  (  0.86,  1.73)   100  -0.01536
AxR  0.49  (  0.20,  2.69)  ( -0.12,  0.91)    27   0.00289
AxN  1.21  (  0.73,  4.46)  (  0.43,  2.13)    46  -0.02755
AxD  1.16  (  0.51, 17.73)  ( -0.30,  2.23)    31   0.02027
AxC  1.80  (  0.85,  9.93)  ( -0.09,  3.28)    35   0.01020
AxQ  0.91  (  0.57,  1.61)  (  0.38,  1.30)    31   0.01439
AxE  0.83  (  0.57,  2.17)  (  0.45,  1.35)    41  -0.03175
AxG  1.12  (  0.74,  2.64)  (  0.55,  1.77)    95  -0.01972
AxH  1.37  (  0.89,  2.58)  (  0.54,  2.01)    95   0.02223
AxI  0.99  (  0.79,  1.40)  (  0.74,  1.28)   203  -0.00775
AxL  0.93  (  0.80,  1.11)  (  0.77,  1.08)   387  -0.00119
AxK  0.73  (  0.39,  2.00)  (  0.14,  1.21)    27   0.01665
AxM  1.60  (  1.16,  2.48)  (  0.98,  2.18)   183   0.01744
AxF  1.20  (  1.03,  1.52)  (  1.00,  1.45)   382  -0.00643
AxP  2.24  (  1.76,  3.45)  (  1.61,  3.03)   139  -0.03102
AxS  0.92  (  0.52,  1.94)  (  0.22,  1.40)    85   0.02318
AxT  0.90  (  0.58,  1.57)  (  0.39,  1.29)   113   0.01796
AxW  1.18  (  0.93,  1.60)  (  0.86,  1.49)   218  -0.00546
AxY  0.85  (  0.56,  1.38)  (  0.42,  1.17)   125   0.00875
AxV  0.84  (  0.60,  1.33)  (  0.50,  1.15)   171  -0.00463
RxR  1.17  (  0.28,   inf)  ( -1.58,  2.33)    11   0.09892
RxN  0.75  (  0.27,   inf)  ( -0.44,  1.49)    10   0.00945
RxD  3.17  (  1.76,  8.51)  (  0.64,  5.18)    30   0.01924
RxC  0.44  (  0.07,   inf)  ( -0.61,  0.87)     3   0.00998
RxQ  2.08  (  1.06, 17.03)  (  0.14,  3.92)    25  -0.00676
RxE  2.17  (  1.17,  6.30)  (  0.36,  3.61)    38   0.00720
RxG  0.60  (  0.27,   inf)  ( -0.07,  1.20)    18  -0.02250
RxH  0.25  (  0.10,   inf)  ( -0.09,  0.49)     6   0.00902
RxI  0.19  (  0.08,   inf)  ( -0.08,  0.39)    14   0.00593
RxL  0.70  (  0.54,  0.99)  (  0.49,  0.90)   103  -0.00237
```

Most TM Helices Have H-Bond

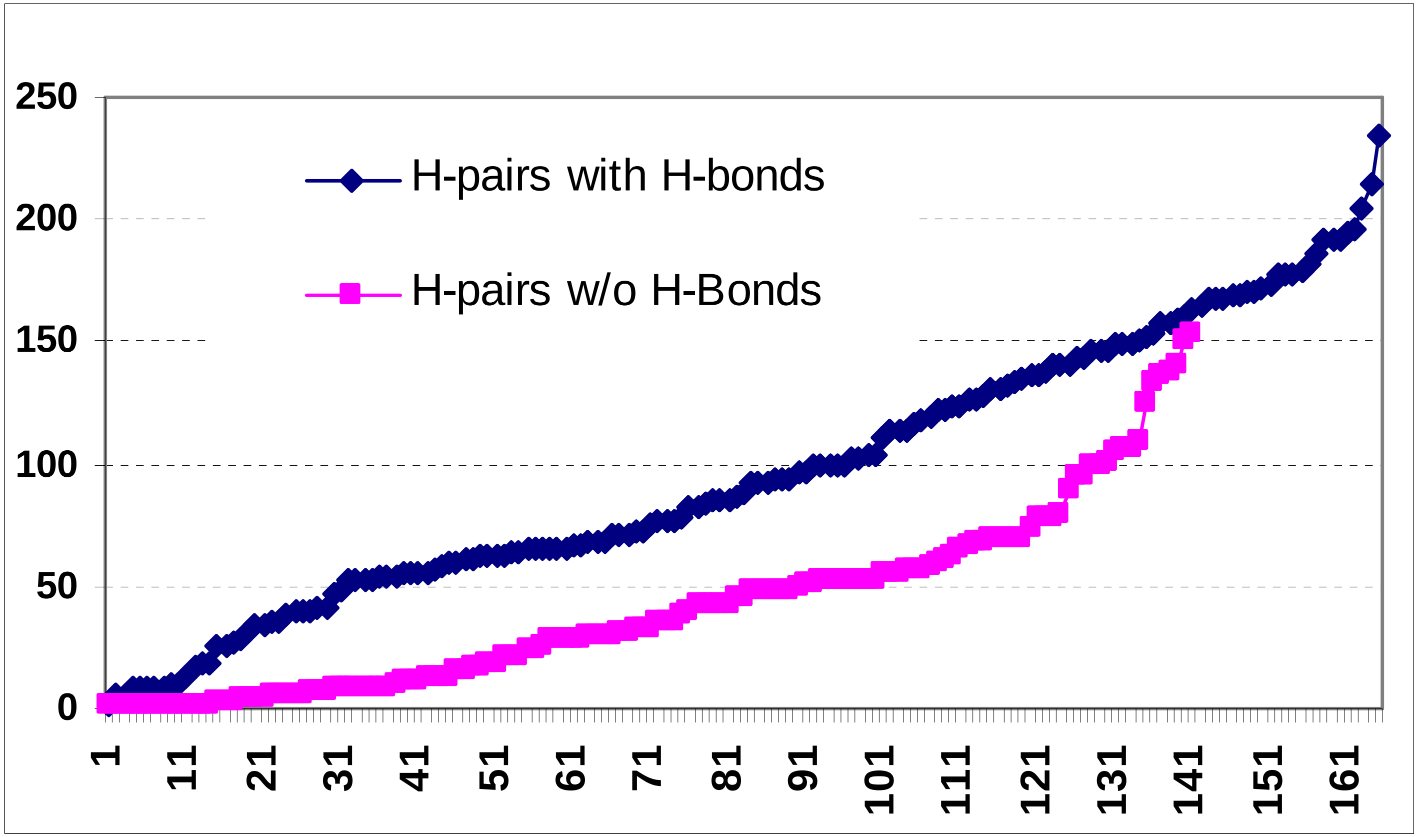
- Exceptions:
  - Low resolution structures:
    - *1be3, 1fum*
  - Weak C_{α}-H—O bond:
    - *1fx8*

Adamian and Liang, Proteins, 2002



Helical Pairs with H-bond Are Packed Tighter

- More interhelical contacts.

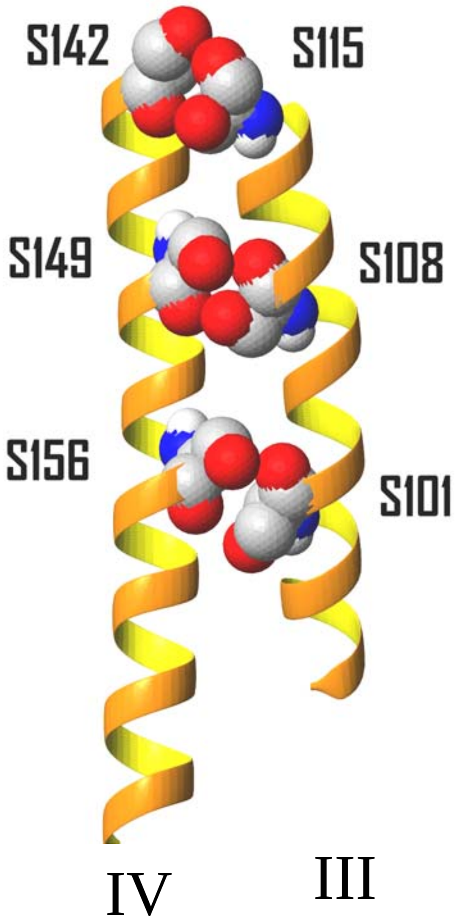


  - Kolmogorov-Smirnov for different distribution: $p = 2.9 \times 10^{-7}$
  - Wilcoxon for different means: $p = 8.1 \times 10^{-20}$

Spatial motif: Serine Zipper

- 3 in cytochrome C oxidase, III & IV
  - — provide tight packing between helices.

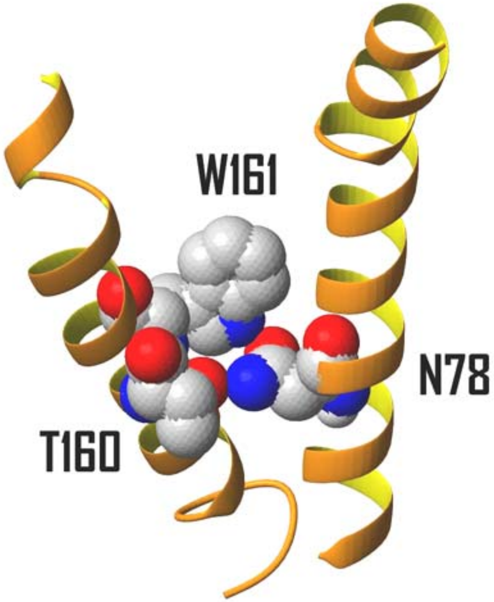
- Heptad motif: similar to leucine zipper.



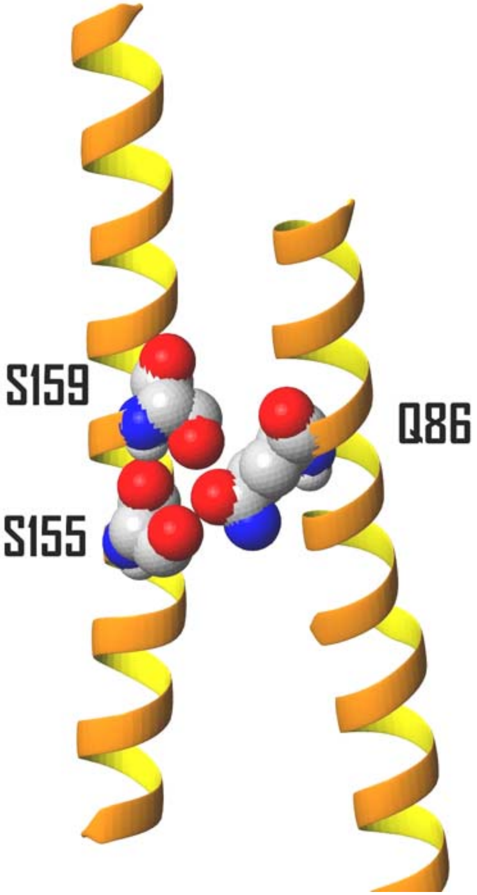
Paracoccus
Cyt C oxidase
1ar1

Spatial motif: Polar Clamps.

- 3 a.a on 2 helices.
- Side chain of N, Q and S, T.
  - N, Q can form 2 H-bonds.
  - Clamped by two H-bonds from 2 a.a. on the other helix:
    (i,i+1), (i,i+3) or *(i,i+4)*
- Very common:
  - 12 out of 13 proteins: except bovine cyt BC1 complex.
- Highly conserved.



Rhodopsin
Helices I & VII
1f88

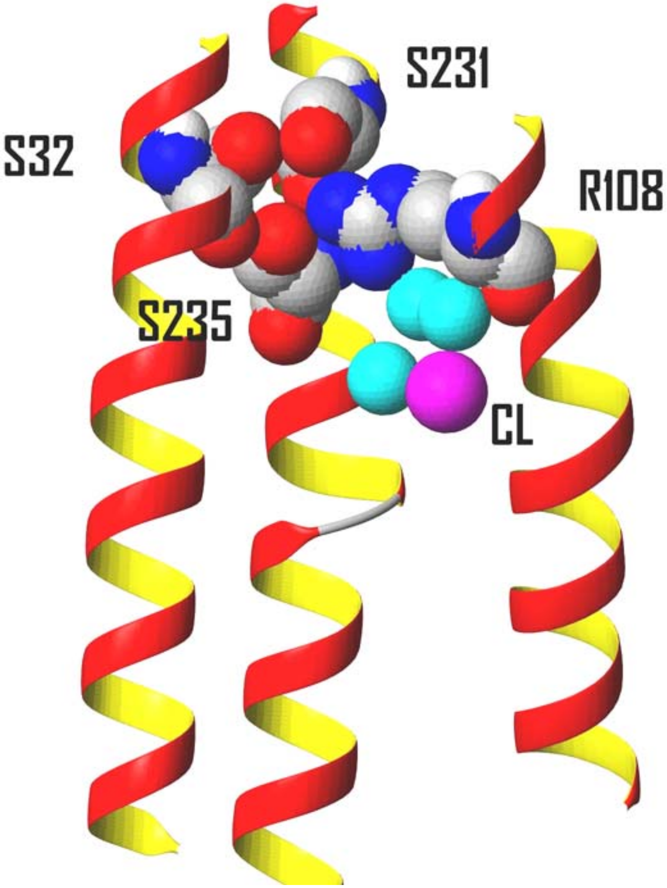


T. thermophilus
Cyt C oxidase
1ehk

- Related to SS4 motif? *(Senes, et al, 00)*

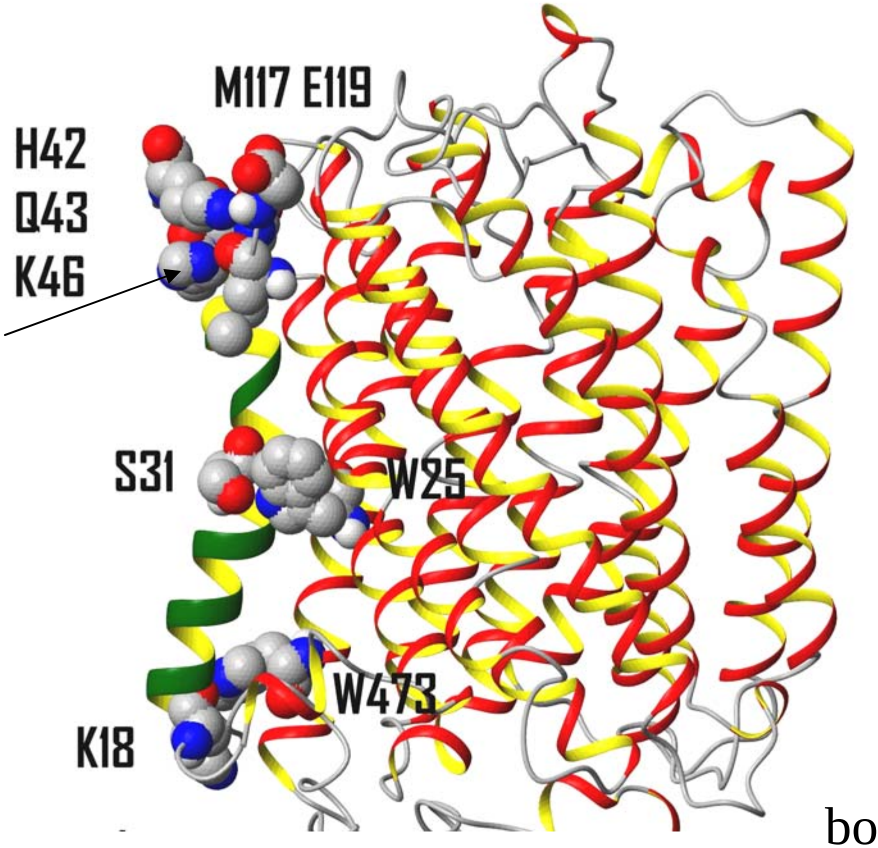
Polar Clamp in Halorhodopsin

- Maybe functionally important.
  - R108K: loss in activity
  - R108Q: no activity, but can be restored by gaunidinium ion. (Rudiger et al, 95)



Polar Clamp in Multisubunit Membrane Protein

- H-bond clusters may determine orientation of single span subunit.



2 Polar Clamps

bovine cyt C oxidase

Helix-Lipid Interactions

(with Larisa Adamian)

Helical Membrane Protein Structure

- How to identify lipid facing surfaces of helices when only sequence is known?

  - What are lipid facing surfaces?
    - Envelope of probe accessible surfaces.
    - Probe radius 1.9 Å modeling methyl group for lipid.
    - Use VOLBL method based on alpha shape from NCSA.

Our approach

- Residue-specific preferential interactions with lipid.
  - Region specific.
  - kPROT propensity scale.
    - Pilpel, Y., Ben-Tal, N., Lancet, D. J. Mol. Biol. 1999(294), 921-935
  - **TMLIP propensity scale.**
    - Adamian, L., Nanda, V., DeGrado, W., Liang, J. (*Proteins*, *2005*).
  - Others: Beuming and Weinstein.

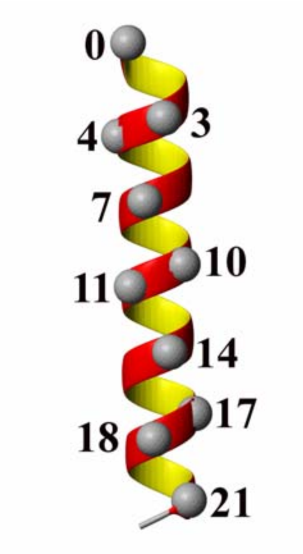
- Core regions are more conserved than lipid-exposed regions.

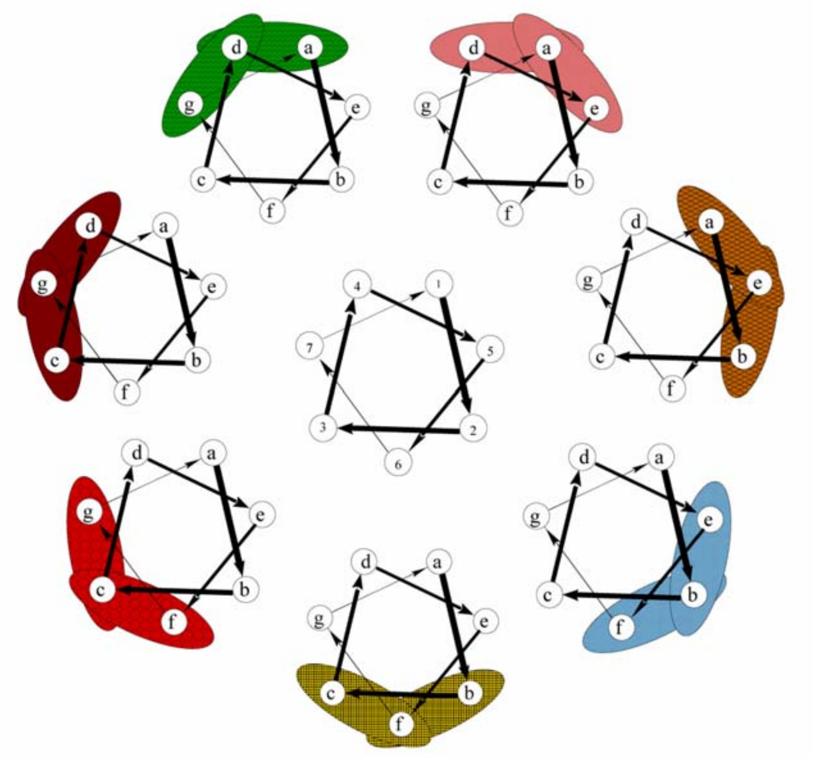
(Steven and Arkin, 2001, Prot Sci)

Cannonical surface of helices

- 7 overlapping helical surfaces centered at each position [*abcdefg*] of the heptad repeat.

Surface	Residues at positions
1	*a-d-e*
2	*b-e-f*
3	*c-f-g*
4	*d-g-a*
5	*e-a-b*
7	*f-c-b*
7	*g-c-d*





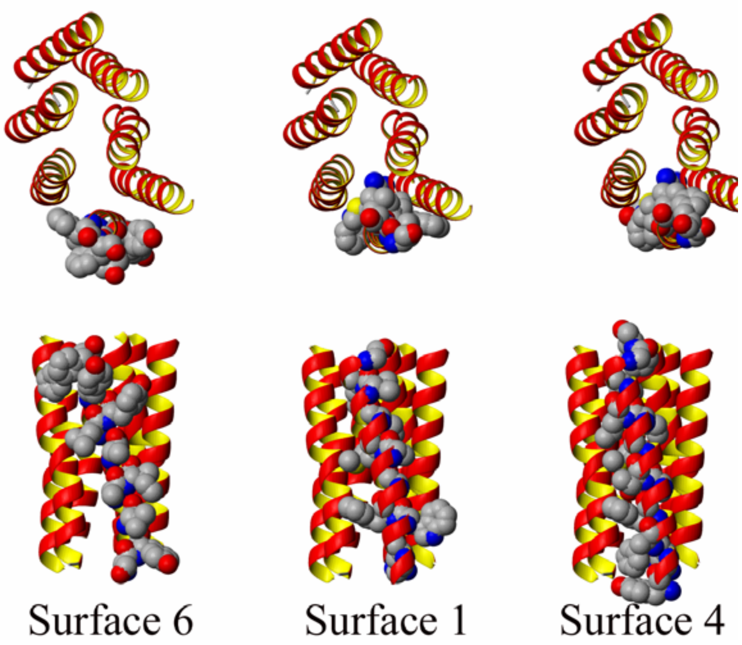
Sequence conservation

- Entropy calculation for each position.
  - Psi-blast: gather sequence, ClustalW: multiple alignment for each helix, Pfaat: gap removal.
  - 35%-90% sequence identity (or 40%-80%).
  - Sequences with functional annotation identical to query sequence.

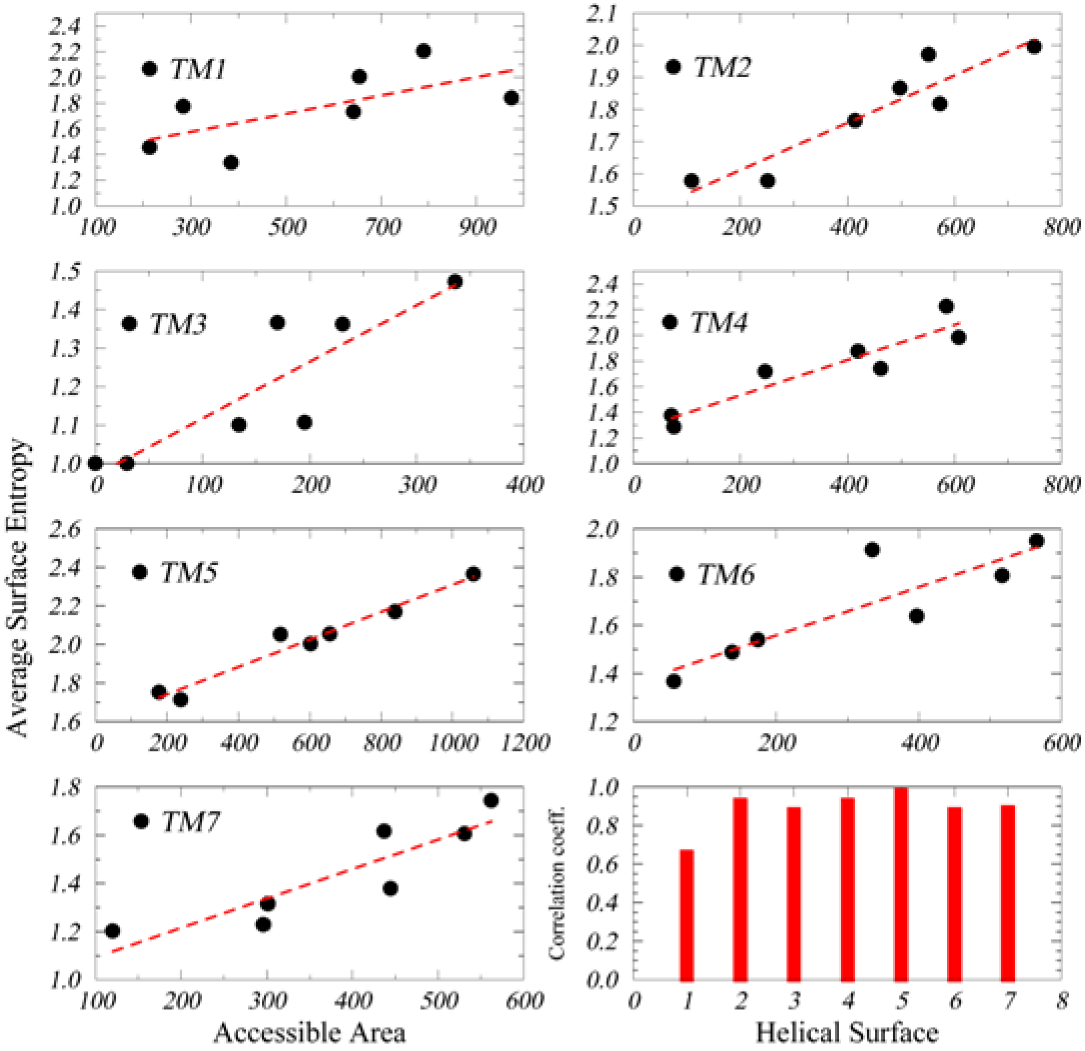
$$E(i) = -\sum_r p_i(r) \ln p_i(r) \quad \text{and} \quad S(i) = e^{E(i)}$$

An example: bR

- Bacteriorhodopsin:1C3W



TM5



LIPS: LIPid Surface prediction

- Predicting lipid exposed surface in membrane proteins.

- Identify helical surface with:
  - Most nonconserved residues.
  - Highest lipid propensity.

- Scoring function:

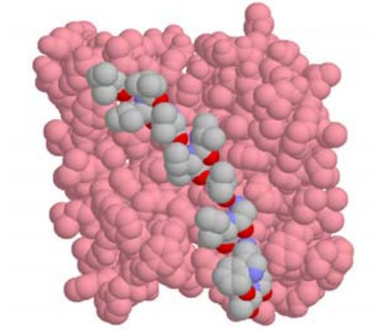
$$H_k = \overline{S_k} + \overline{L}_{TMLIP,k}$$

$\overline{S_k}$: Average positional entropy score for face k.

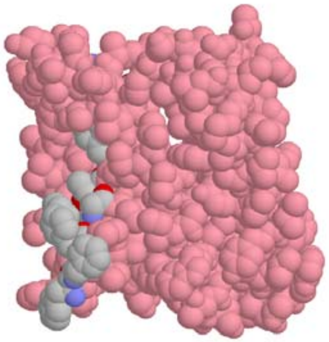
$\overline{L}_{TMLIP,k}$: Average positional lipid propensity score for face k. Obtained from leave - one out data.

Prediction Results: Succinate dehydrogenase

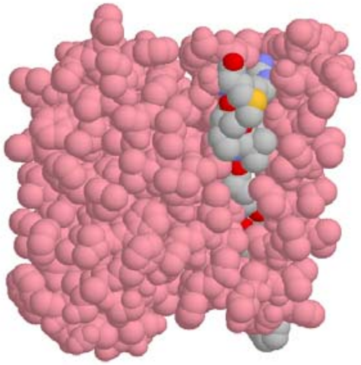
- Helices 2 and 5 cross TM bundle (1nek)
  - With lipid-facing residues on two different sides of the bundle
- Prediction by LIPS shown in gray.



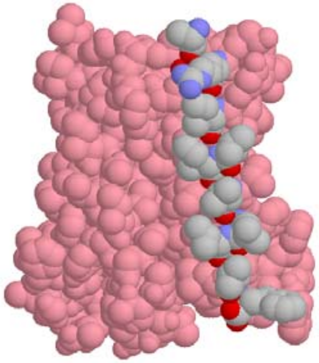
TM1



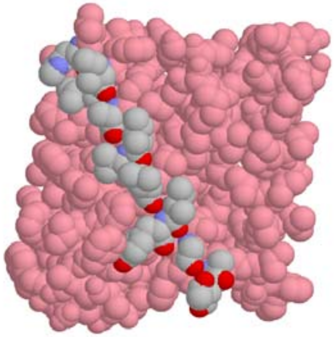
TM2: “front”



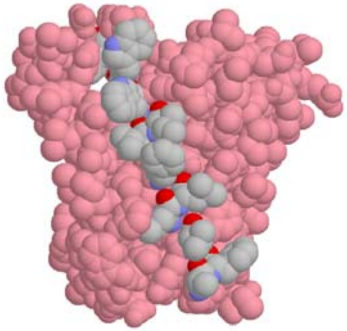
TM2: “back”



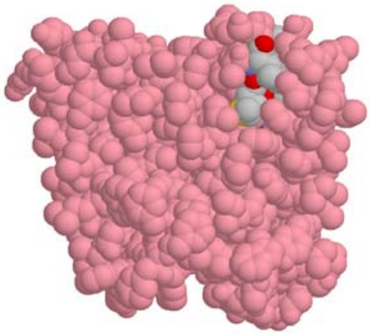
TM3



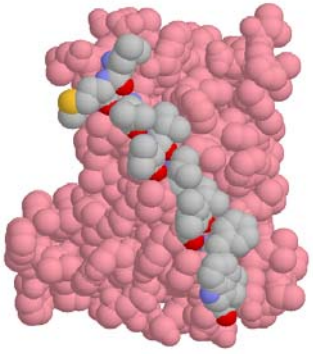
TM4



TM5: “front”

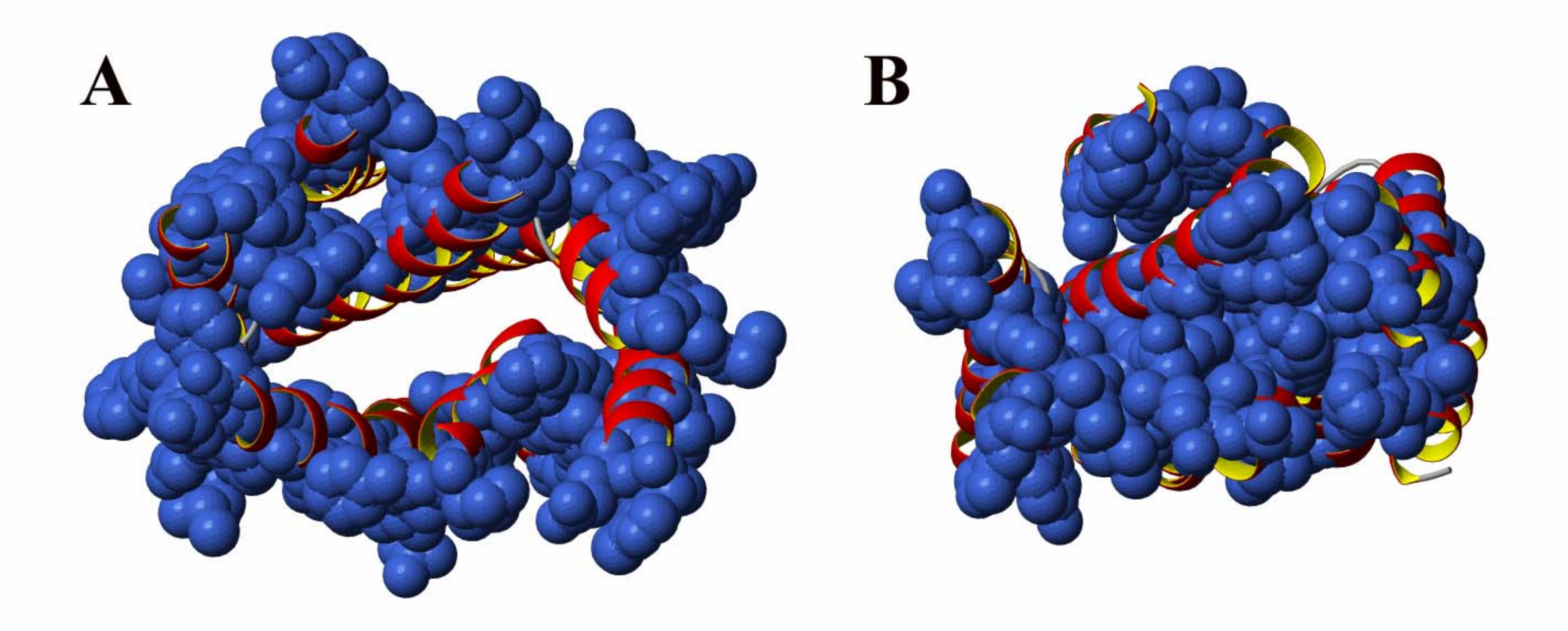


TM5: “back”



TM6

Prediction Results: Bovine rhodopsin



A. Lipid facing residues B. Buried internal residues

Prediction summary

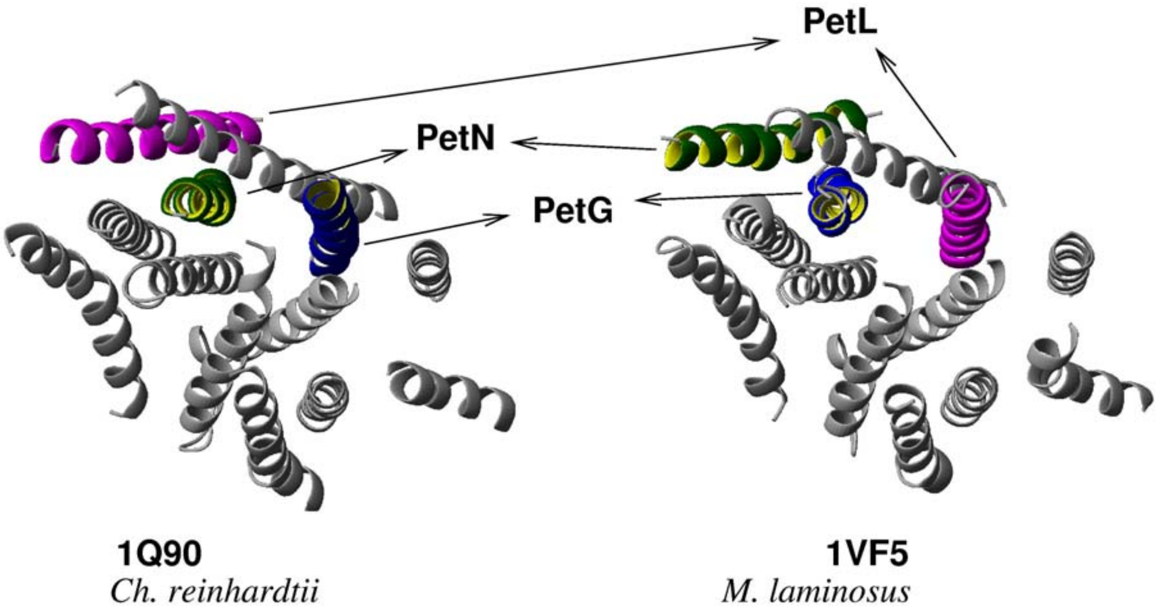
Protein	Total	Lipophilicit	Lipophilicit	Average	Average	LP+AE	LP+AE
	Number of	LP	LP	Entropy AE	Entropy AE		
	Helices	max	min	max	min	max	min
1C3W	7	3	3	7	7	7	7
1EUL	9	2	4	9	6	9	7
1FX8	6	2	1	6	6	6	6
1IWG	12	5	3	8	8	10	8
1J4N	6	4	1	4	5	6	6
1KPL	12	6	7	8	10	10	10
1KQF	5	1	2	4	2	3	2
1L9H	7	3	3	7	7	7	7
1M3X	11	5	4	9	6	10	9
1NEK	6	4	5	6	5	6	5
1OCR	25	15	8	20	11	23	12
1OKC	6	3	2	4	3	6	4
1PV6	12	6	4	8	4	9	5
1PW4	12	7	6	7	7	11	8
1Q16	5	3	5	5	4	5	5
1Q90	10	7	6	8	6	9	9
1RH5	8	5	1	7	3	8	3
Total	158	81	65	127	100	145	113
%	100	51.3	41.1	80.4	63.3	91.8	71.5

All are results from leave-one-out tests.

145 out of 158 (~92%) TM helices from 16 MPs with adequate homologs

Error detection in membrane protein structure

- Transmembrane domains of cytochrome b6f complexes from *Ch. reinhardtii* (pdb:1q90) and *M. laminosus* (pdb:1vf5).
- Assignment of TM helices are very similar, but the assignment of subunits PetG, PetL, and PetN is inconsistent.
- 1vf5 likely wrong: lipid-exposed conserved faces, bad H-bonds



Summary

- Helix-helix interactions.
  - H-bond, and spatial patterns.
- Helix-lipid interface prediction.
  - Buried helices.
  - Lipid facing surface.

Collaborators

- Ronald Jackups, Larisa Adamian/UIC
- Bill DeGrado, Vikas Nanda / U Penn.

Acknowledgement

- NSF CAREER DBI 0133856 and DBI 0078270
- NIH GM68958