Advances in computational protein design

Applying computation to protein design

- how efficient is it?
- how reliable is it?

Evolving roles of computation in protein design

- molecular modeling and energy calculation
- side chain prediction
- redesign of a protein
  - sometimes limited in scope
  - expert interpretation common
- more challenging designs that rely exclusively on computation

Feasibility study

- redesigning the core of an existing protein
- large scale design that depends entirely on computation
- FSD-1 : first example of a protein designed entirely based on computation
- Top7 : much more ambitious, new topology not seen in nature

Function-oriented design

- Design specific interactions
  - modulate specificity and affinity
  - calmodulin-peptide
  - ligand-receptor problems
  - integrin
- Catalyst (enzyme) design
  - introducing novel catalytic activity
  - “protozyme”, retro aldolase
  - combining computation with library screening
  - DNA endonuclease
- Evaluate thermodynamics
  - computational ala scanning—predicting the impact of ala substitution
- Negative design
  - both structural and functional

Core packing

Many designed proteins lack a well-defined, unique, tertiary structure despite their high thermal stability

If core residues do not pack specifically, these proteins behave as if they are molten globules

Can we improve core packing computationally?

Repacking of Cores (ROC)

Genetic algorithm-based program to introduce a large number core residue substitutions to explore alternative packing

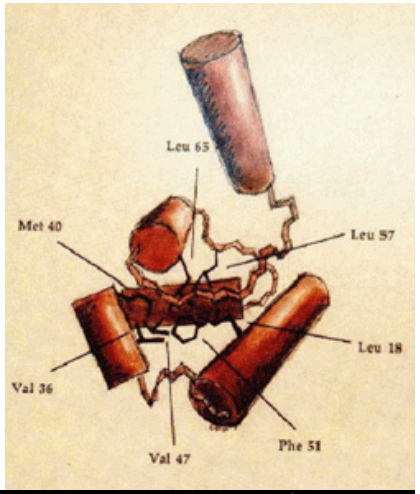
- require prediction of side chain structure, core sequence, relative stabilities in natural proteins
- custom rotamer library—e.g. different rotamer set for each buried position
- search for global optimum
- Desjarlais and Handel, Protein Sci, 4, 2006 (1995)

Repacked Proteins

Computational repacking of lambda repressor produced sequences similar to those found by Lim and Sauer, 1989

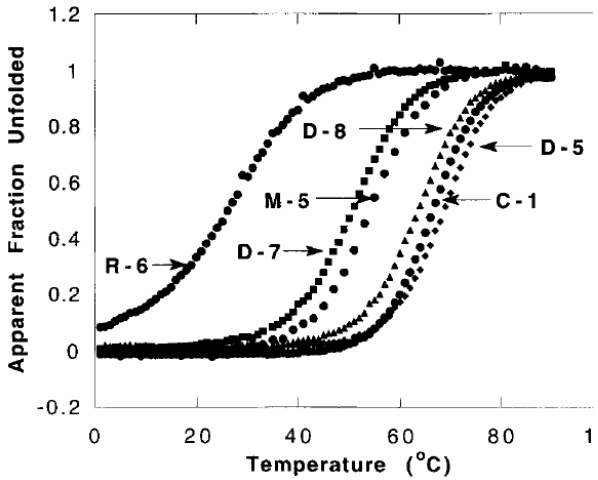
Bacteriophage 434 Cro

- keep non-core side chains and mutated core residues
  - » core residues are easier to re-design
- control: six randomly generated isosteric substitutions
- “minimalist” core: mostly leucine residues

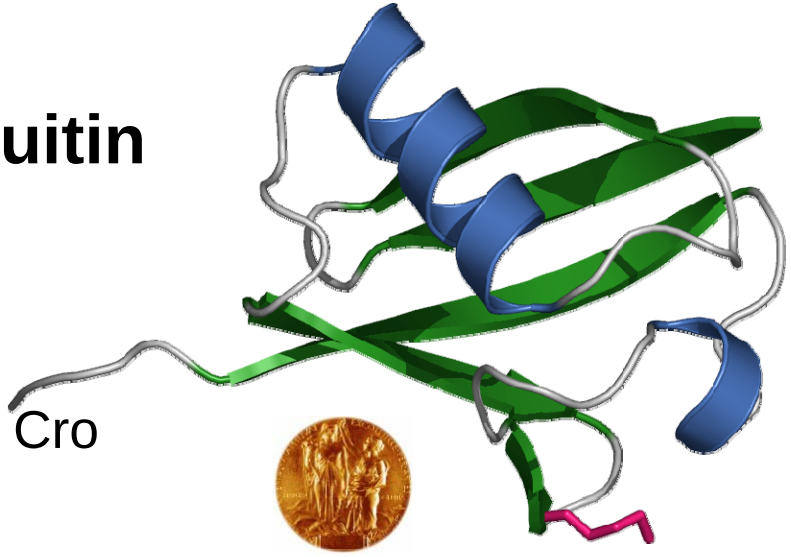


	2	6	13	20	26	31	34	45	48	52	54	58	59	E_{tot}	E_{s-b}	E_{s-s}	ΔVol	T_m
C-1	L	L	L	L	V	I	I	L	I	L	V	W	L	-64.3	-87.6	-15.0	+36	56
D-5	I	I	L	L	I	I	L	L	I	L	V	W	L	-68.8	-87.0	-15.8	+64	60
D-7	I	F	L	V	L	V	I	L	L	L	V	W	L	-66.2	-85.6	-17.0	+44	17
D-8	F	I	L	L	L	V	L	I	L	L	V	W	L	-67.0	-81.3	-17.3	+70	50
M-5	L	L	L	L	L	L	L	L	L	L	V	W	L	-59.8	-88.1	-13.3	+59	33
R-6	I	L	I	L	V	L	I	I	L	L	V	W	L	-52.8	-80.6	-11.7	+37	-

Desjarlais and Handel, Protein Sci, 4, 2006 (1995)



Repacked ubiquitin



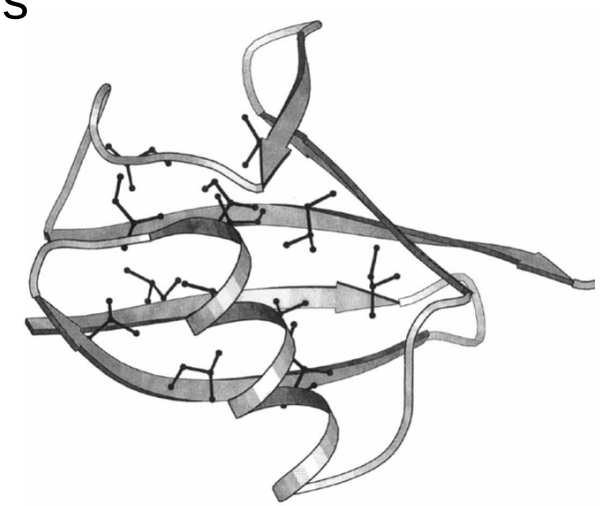
2004 Chemistry

Can ROC redesign beta sheet proteins?

Ubiquitin has a more complex topology than Cro

- involved in proteolytic degradation
- high initial stability—more engineerable
- small (76 residues), soluble, good NMR spectra
- structural, dynamic, kinetic folding data available for WT

Choose a mutant with a large number of substitutions to achieve a dramatically different core packing

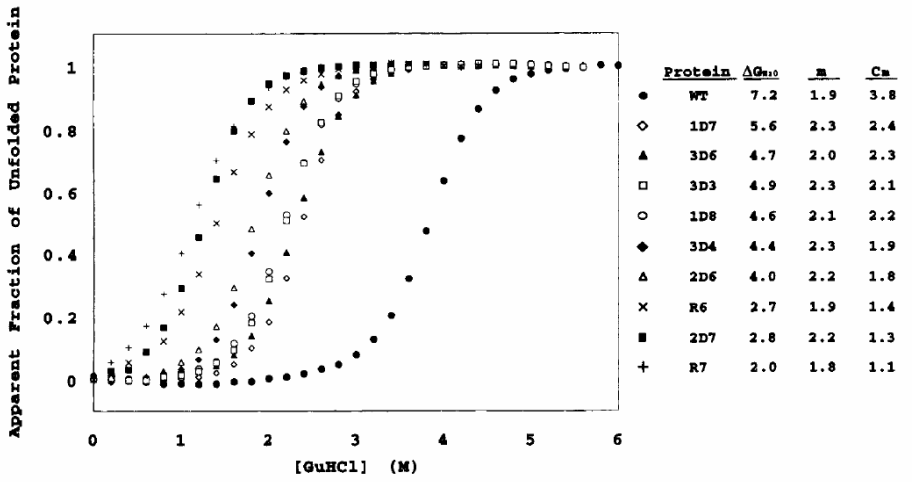


Optimizing the potential function parameters and rotamer library maximizes the correlation between thermodynamic data and predicted stabilities

Lazar et al, Protein Sci 6, 1167 (1997)

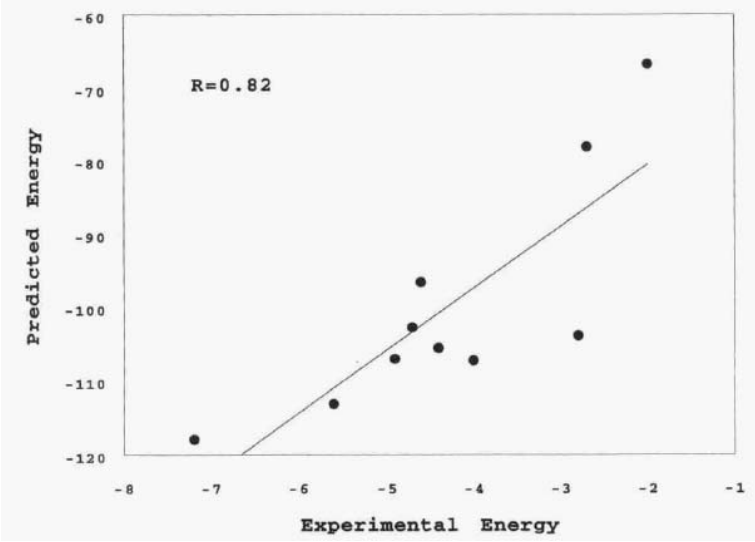
Protein	Energy	ΔV	Class	Residue													
				3 s	5 s	13 s	15 s	17 s	23 h	26 h	30 h	43 s	50 s	56 c	61 c	67 s	69 s
WT	-117.9	-	-	I	V	I	L	V	I	V	I	L	L	L	I	L	L
1D8	-96.5	-3	I	**L**	V	**L**	**V**	**L**	**V**	**L**	I	L	L	L	**L**	**I**	L
1D7	-113.1	+8	I	**V**	**L**	**V**	**I**	V	**V**	**F**	I	L	L	L	I	**I**	L
2D6	-107.2	+25	I	**L**	**I**	**V**	L	V	I	**L**	**L**	L	L	L	I	L	**I**
2D7	-103.8	+60	II	**L**	**L**	**V**	L	V	I	**I**	**L**	**F**	L	L	I	L	**I**
3D6	-102.7	0	I	**L**	V	I	**I**	**I**	**V**	V	I	L	L	L	**L**	**I**	L
3D4	-105.5	-29	I	**L**	V	I	L	**L**	**V**	V	I	L	L	L	**V**	L	L
3D3	-107.0	-2	I	**L**	V	I	L	**L**	**V**	V	I	L	L	L	I	L	L
R7	-66.7	0	II	I	**I**	**L**	**V**	V	I	V	**L**	L	**I**	**I**	**L**	L	L
R6	-72.5	+30	II	I	V	I	**I**	**I**	I	V	**L**	**I**	**I**	**I**	I	L	L

prediction



Protein	ΔG_{H_2O}	m	C_m
WT	7.2	1.9	3.8
1D7	5.6	2.3	2.4
3D6	4.7	2.0	2.3
3D3	4.9	2.3	2.1
1D8	4.6	2.1	2.2
3D4	4.4	2.3	1.9
2D6	4.0	2.2	1.8
R6	2.7	1.9	1.4
2D7	2.8	2.2	1.3
R7	2.0	1.8	1.1

thermodynamic measurement



correlation

Protein design automation

Design all parts of a protein, including non-core residues

- core residues interact mostly through van der Waals contact
- surface residues have much greater degrees of freedom
- solvation effects must be accounted for—electrostatic interaction is dampened

Must be able to design novel protein objectively

- algorithm based on physicochemical potential function
- mathematical description of stereochemical constraints
- use knowledge obtained from manual design and protein folding
- fully automated, unbiased, quantitative approach that can be rigorously tested

FSD-1

ORBIT

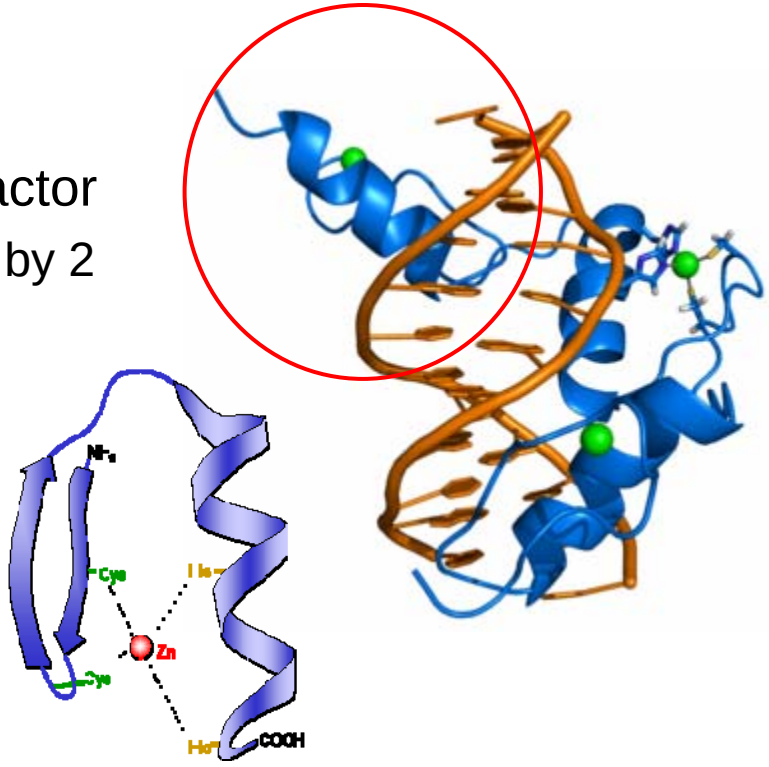
Dead-end elimination-based program to search through a large combinations of rotamer states

Start with backbone fold and search for sequence to stabilize target structure

Iterative optimization of solvation parameters using experimental data and simulations

Zinc finger protein (Zif268) is a transcription factor

- small domain stabilized by Zn++ coordinated by 2 cys and 2 his
- beta-beta-alpha motif has been engineered by hand—ref. Struthers et al, Science 271, 342 (1996)



Dahiyat and Mayo, Science 278, 82 (1997)

Core: mutate to A, V, L, I, F, Y, W

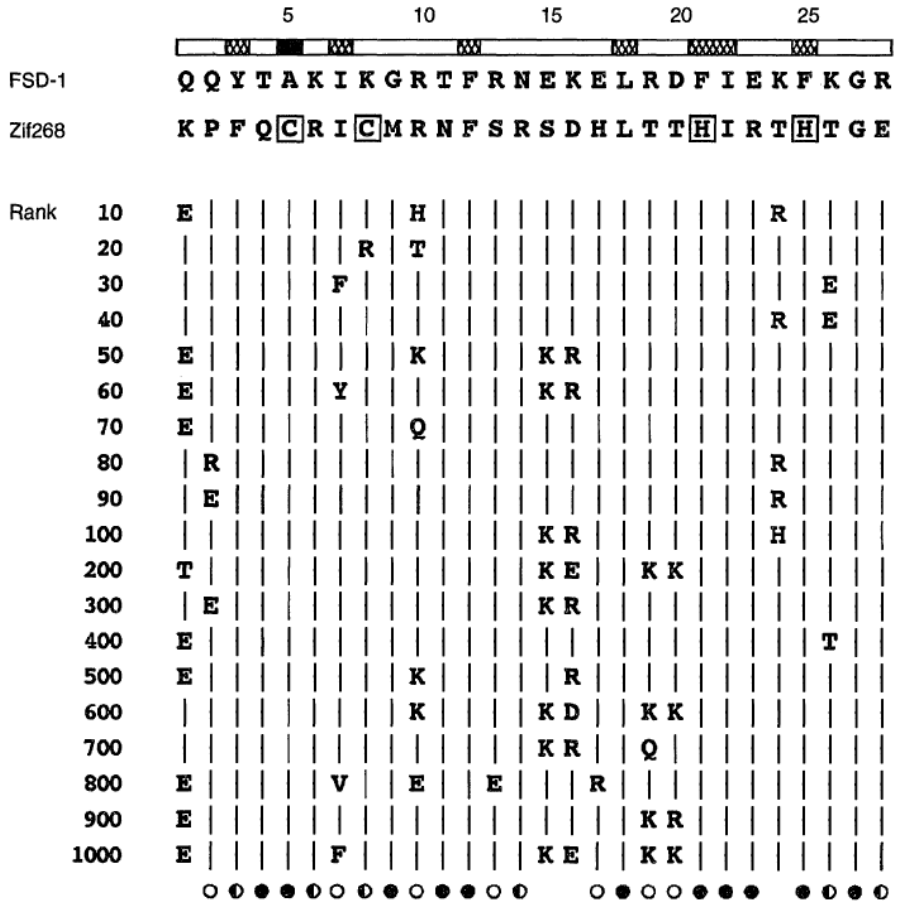
Surface: mutate to A, S, T, H, D, N, E, Q, K, R

Boundary: mutate to sum of core and surface

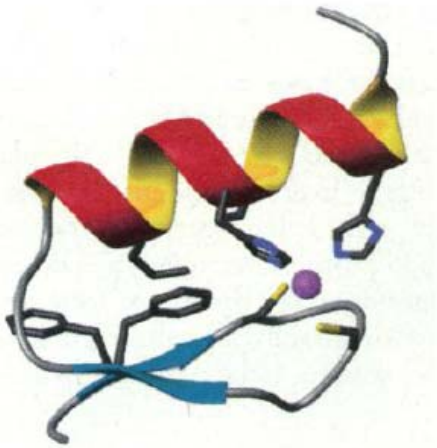
Res #9, 27: phi > 0° ➔ Gly to minimize backbone strain

Screen 1.9 x 10^27 possible sequences or 1.1 x 10^62 rotamers

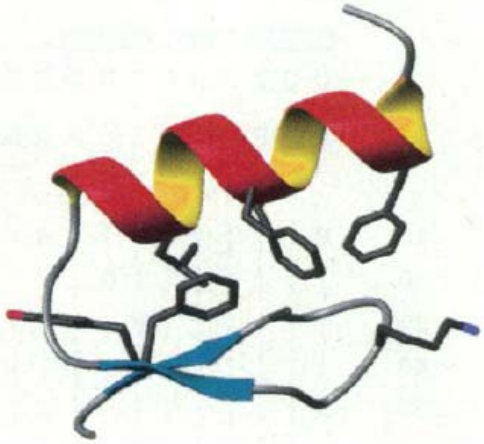
Determine the structure of FSD-1 by NMR



wt



fsd-1



New topology design

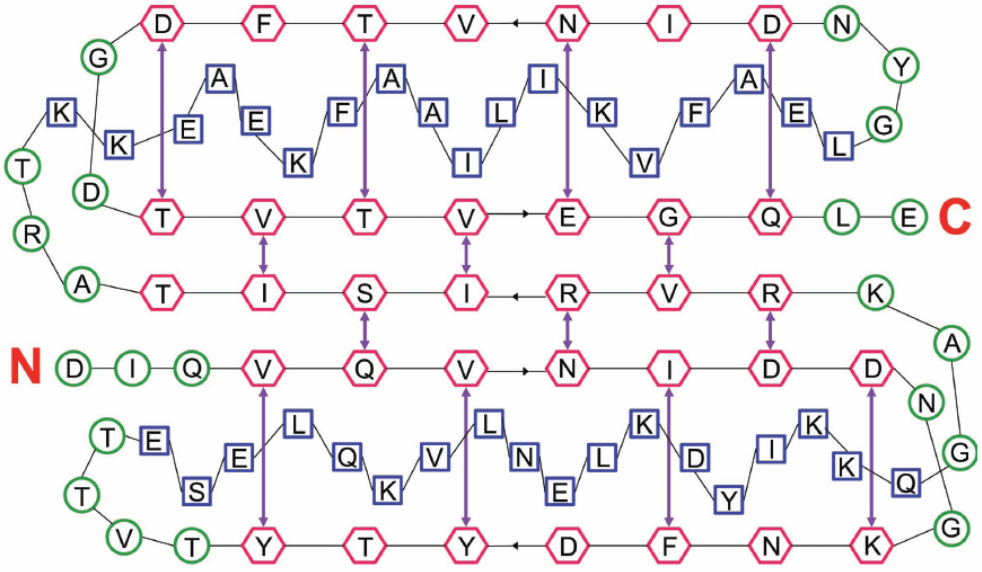
Only a limited amount of structural diversity is found among native protein

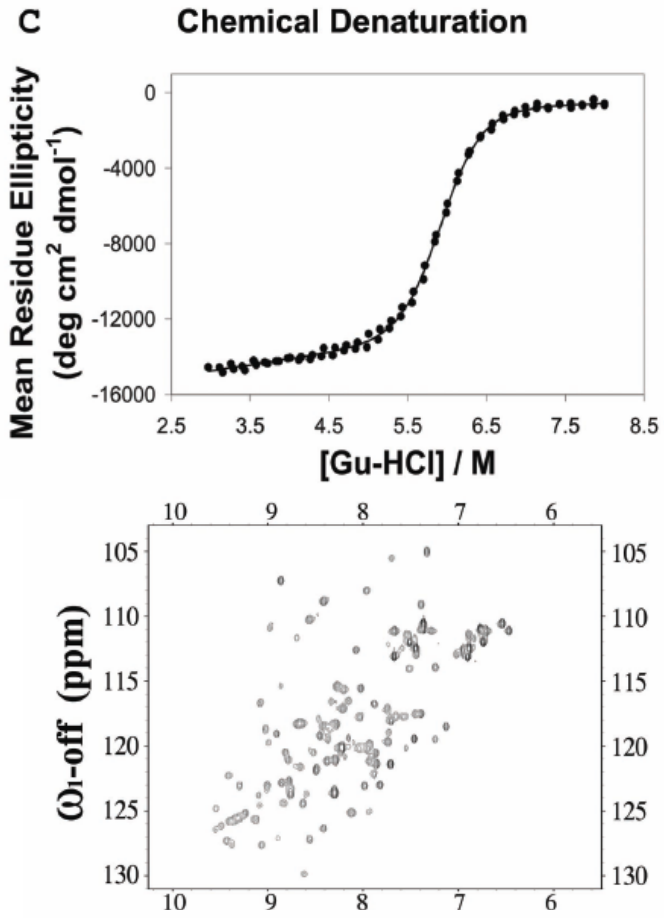
- are other structures disallowed for physical reasons?
- can we design a protein with a new topology?
- unlikely that an arbitrarily chosen structure will be designable—need to simultaneously search both sequence and structure spaces

RosettaDesign

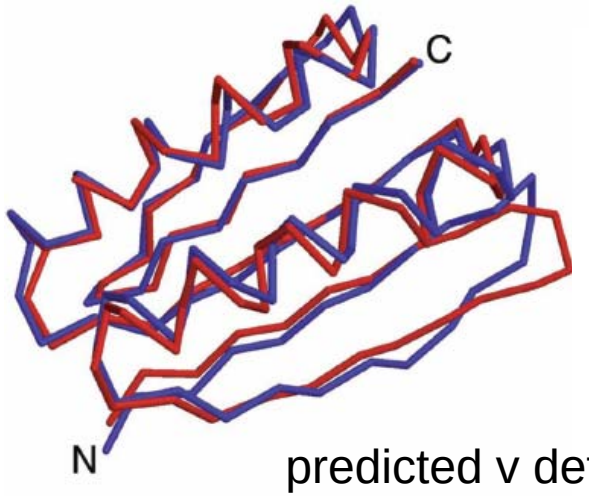
- Generate a target structure by grafting 3 – 9 residue fragments from PDB
- Five stranded sheet and two helices (Top7)
- Design a starting sequence by searching through > 10^{186} combinations
- Iterate between **Monte Carlo**-based sequence optimization for a fixed backbone conformation and gradient-based optimization of the backbone
- 15 cycles of sequence design and backbone optimization
- Parameterization of the atomic radii to dampen Lennard-Jones repulsion

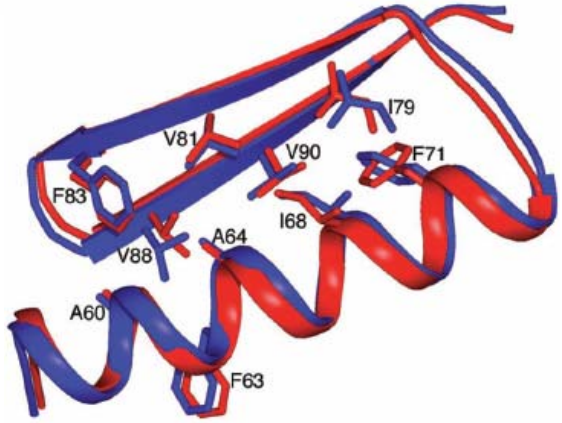
Kuhlman et al, Science 302, 1363 (2003)





betanova





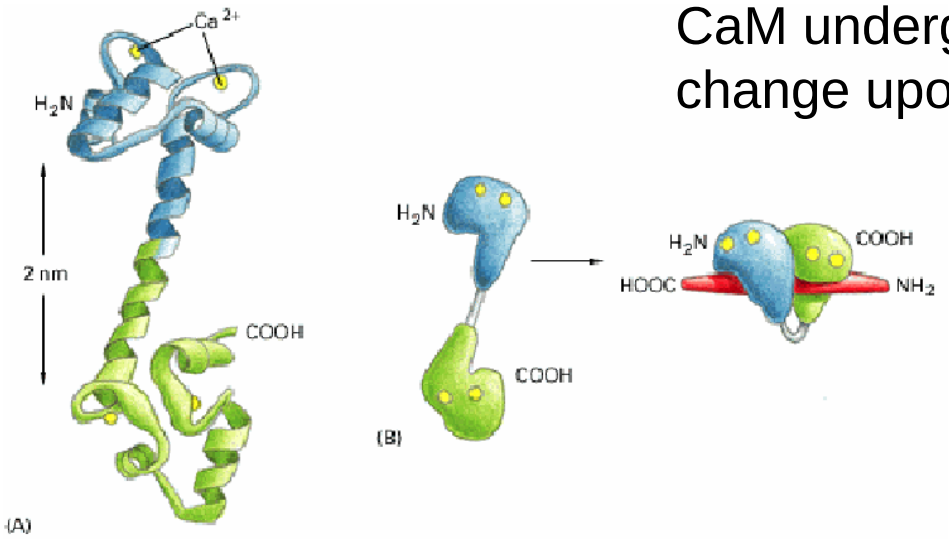
predicted v determined

Designing specificity

Can computation be used to identify and engineer binding specificity in proteins?

Designing specificity is equivalent to identifying a combination of amino acids at the interface that would interact with one another stably

Calmodulin (CaM) is a ~150 residue, Ca++ binding protein that controls many biochemical processes in cells

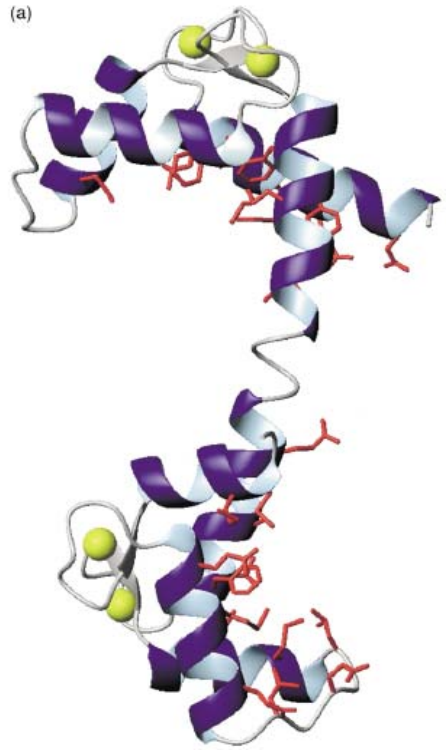


CaM undergoes a large conformational change upon binding its helix ligand

Yet CaM binds a broad spectrum of sequences

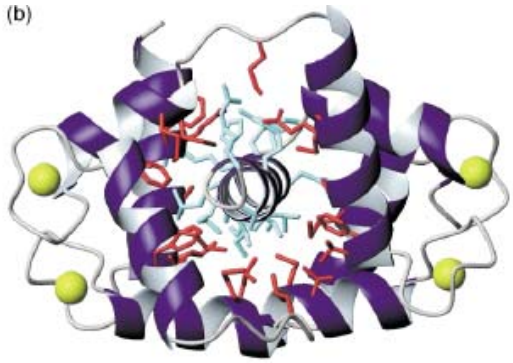
Table 1. Sequence alignment of CaM target peptides

smMLCK						A	R	R	K	W	Q	K	T	G	H	A	V	R	A	I	G	R	L	S	S					
skMLCK								K	R	R	W	K	K	N	F	I	A	V	S	A	A	N	R	F	K	K	I	S	S	S	G	A
Spectrin							K	T	A	S	P	W	K	S	A	R	L	M	V	H	T	V	A	T	F	N	S	I	K	E		
Melittin			Q	Q	R	K	R	K	I	W	S	I	L	A	P	L	G	T	T	L	V	K	L	V	A	G	I	G			
Peptide 1									L	K	W	K	K	L	L	K	L	L	K	K	L	L	K	L	G					
CaMKK							R	F	P	N	G	F	R	K	R	H	G	M	A	K	V	L	I	L	T	D	L	R	P	I	R	R	V
CaMKII	L	K	K	F	N	A	R	R	K	L	K	G	A	I	L	T	T	M	L	A	T	R	N	F	S							



Optimize the interface between CaM and smMLCK in the complex structure

- 24 buried CaM residues within 4 Å of the ligand were optimized
- allow A, V, L, I, W, F, Y, M, E (abundant in CaM interface)
- residues in smMLCK were allowed to change conformation



Shifman and Mayo, JMB 323, 417 (2002)

- introduce 8 mutations at the interface
- 3 Met (responsible for promiscuity) mutated to other residues
- binding affinity for the target ligand increased from 1.8 nM to 1.3 nM
- affinity to other target peptides decreased by 1.5 to 86 fold

smMLCK fluorescence

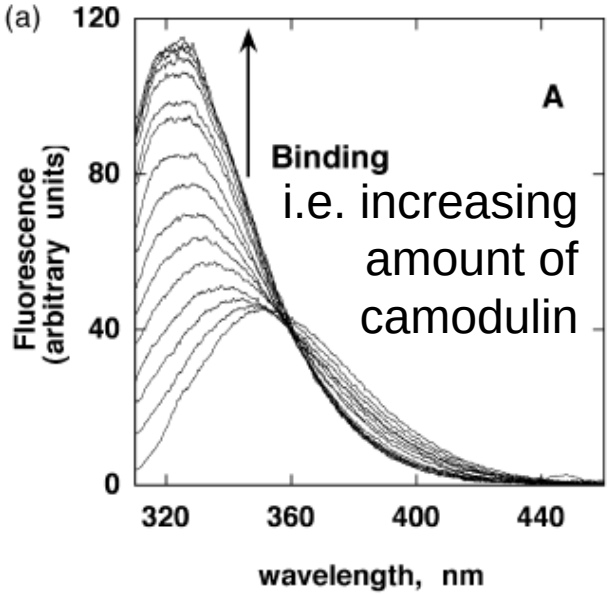


Table 2. Computationally designed CaM mutant

	Designed positions																							
	11	12	15	18	19	32	36	39	51	55	68	71	72	76	84	88	91	92	108	109	112	124	144	145
WT	E	F	A	L	F	L	M	L	M	V	F	M	M	M	E	A	V	F	L	M	L	M	M	M
CaM_8	L	Y	–	–	–	–	–	–	–	I	–	–	–	E	Y	–	I	–	–	L	–	–	–	I

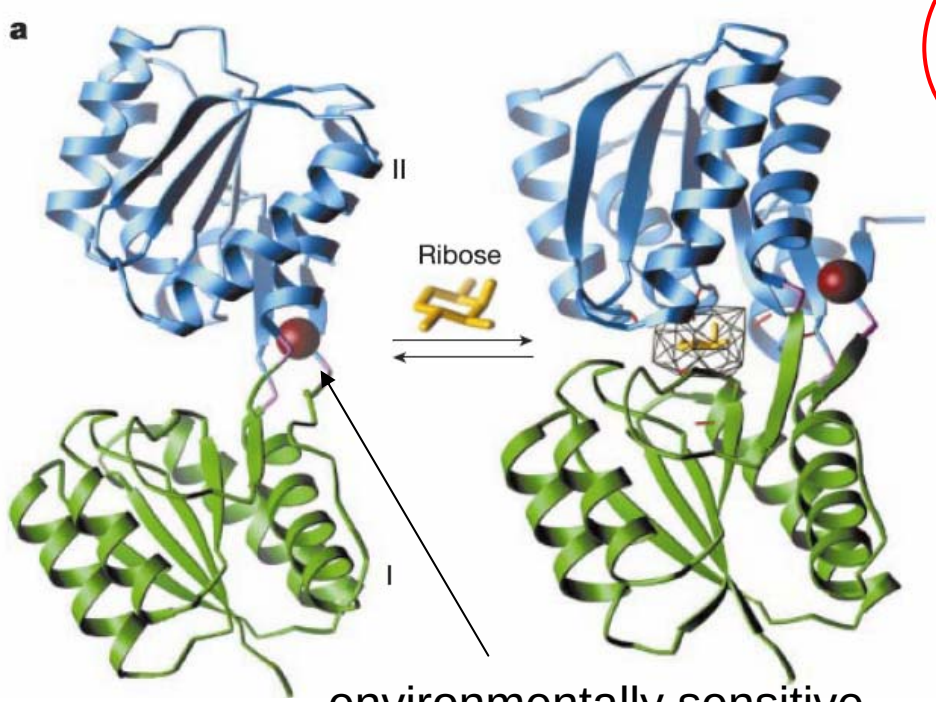
Table 3. Binding affinities of selected targets to WT and redesigned CaM

	Target peptides						
	smMLCK	skMLCK	Spectrin	Melittin	Peptide I	CaMKII	CaMKK
CaM WT	1.8 ± 1.3	3.3 ± 0.8	3.3 ± 1.5	28 ± 5.0	1.7 ± 0.8	5.1 ± 1.5	1.0 ± 3.0
CaM_8	1.3 ± 0.9	4.9 ± 1.2	16 ± 6.0	54 ± 18	147 ± 48	54 ± 20	32 ± 13
α^a	1.0	2.1	6.7	2.6	120	15	44

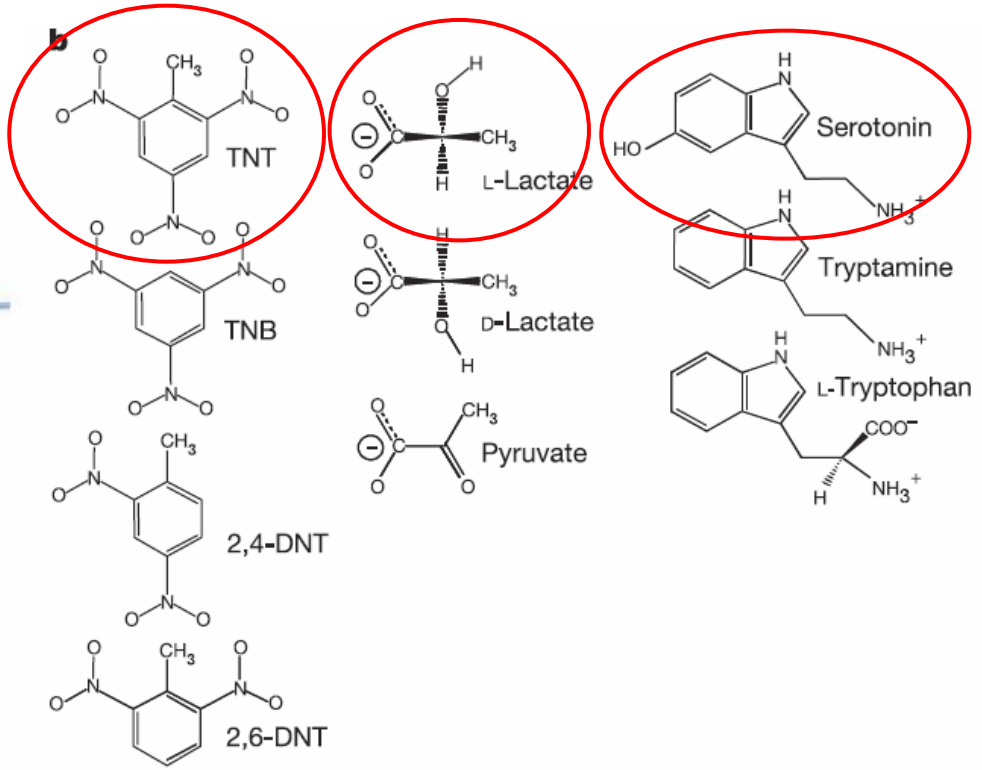
Receptor design

Structure-based computational method to introduce specificity and high affinity for novel ligands into five periplasmic binding proteins of E. coli

- designed receptor may function as a biosensor
- ligand with drastically different chemical properties
- use structural information to optimize short range interactions (“lock and key”)
- discriminate against decoys



environmentally sensitive fluorescent dye for readout

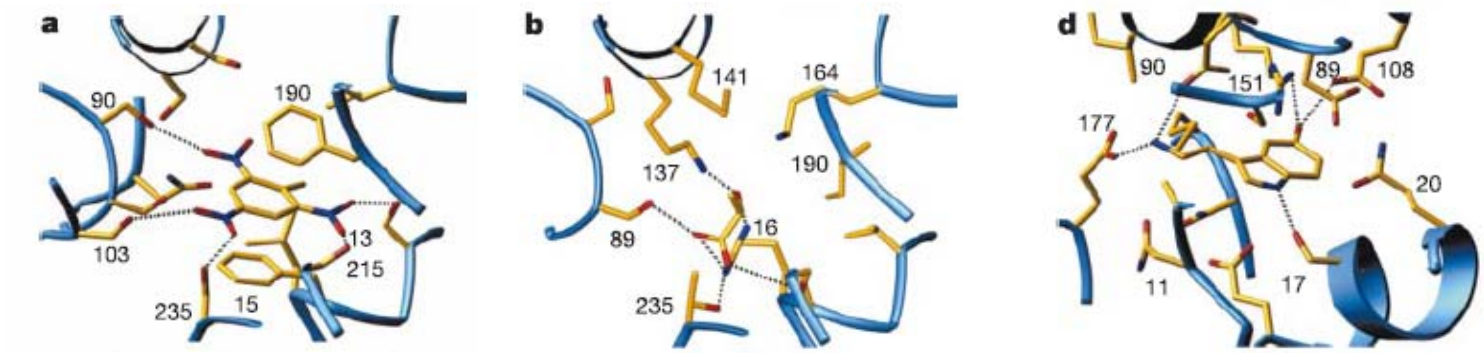


Looger et al, Nature 423, 185 (2003)

- designed ligands are chemically distinct from wt cognate ligands (ribose, glucose, arabinose, his, gln)
- assess the roles of molecular shape, chirality, functional groups (e.g. nitro group of TNT, hydroxyl, carboxylate, amine), polarity (polar, aliphatic, aromatic), charge (neutral, anionic, cationic), solubility
- design complementary surfaces using DEE

- introduce 5 – 17 amino acid substitutions

RBP			9_I	13_I	15_I	16_I	64_I	89_I	90_I	103_H	132_H	137_H	141_{II}	164_{II}	190_{II}	214_{II}	215_{II}	235_{II}
		Wild-type	S	N	F	F	N	D	R	S	I	A	R	F	N	F	D	Q
	TNT	R1 (I)	S	N	A	N	S	S	R				R	S	N		A	S
		R2 (I)	S	I	A	N	N	A	D		K		A	N	N	K	A	N
		R3 (A)		S	F	L		S	S	S		S	S	I	F		S	S
	Lac	R1 (A)		V	A	R		S	S	S		K	M	K	I		S	T



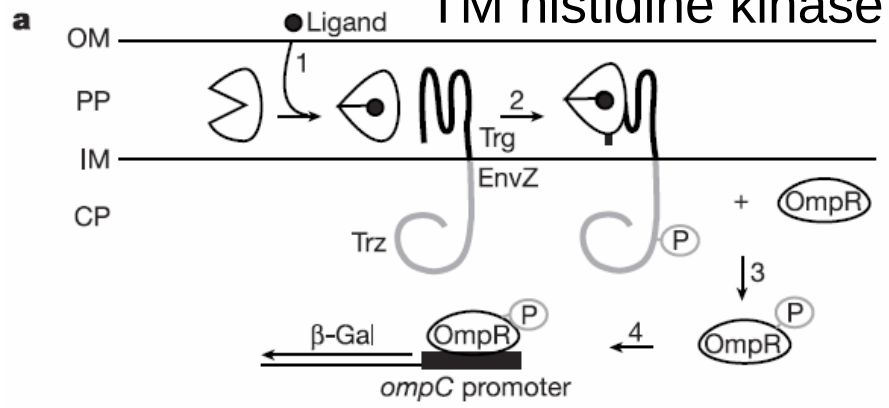
TNT lactate serotonin

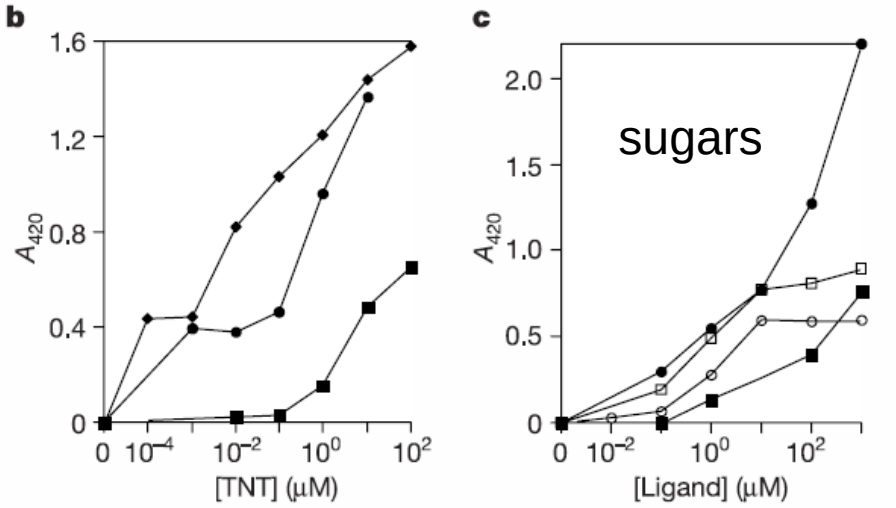
- high affinity and specificity for target molecules
  - kd ~ 2 nM for TNT and weak interaction with decoys (0.1 – 1.5 uM)

Table 2 **Affinities of the designed receptors for target ligands and analogues**

		K_d (μM)			
Target	Receptor	TNT	TNB	2,4-DNT	2,6-DNT
TNT	RBP.R1	0.34	1.0	5.0	5.4
	RBP.R2	1.6	3.8	5.3	4.9
	RBP.R3	0.002	0.1	8.4	15
	ABP.A1	1,400	600	>10,000	>10,000
	ABP.A2	400	500	2,000	4,000
	HBP.H1	220	1,000	>10,000	>10,000
		L-Lac		D-Lac	Pyr
L-Lactate	GBP.G1	2.8		205	255
	GBP.G2	2.1		55	115
	HBP.H1	1.8		40	50
	HBP.H2	12.2		30	48
	QBP.Q1	9,500		>100,000	>100,000
	QBP.Q2	300		>100,000	>100,000
	QBP.Q3	25,000		>100,000	>100,000
	ABP.A1	160		>100,000	>100,000
	ABP.A2	20,000		>100,000	>100,000
	RBP.R1	7.4		40	40
		Stn		Trp	Trm
Serotonin	ABP.A1	50		660	900
	ABP.A2	4.7		65	90

TM histidine kinase



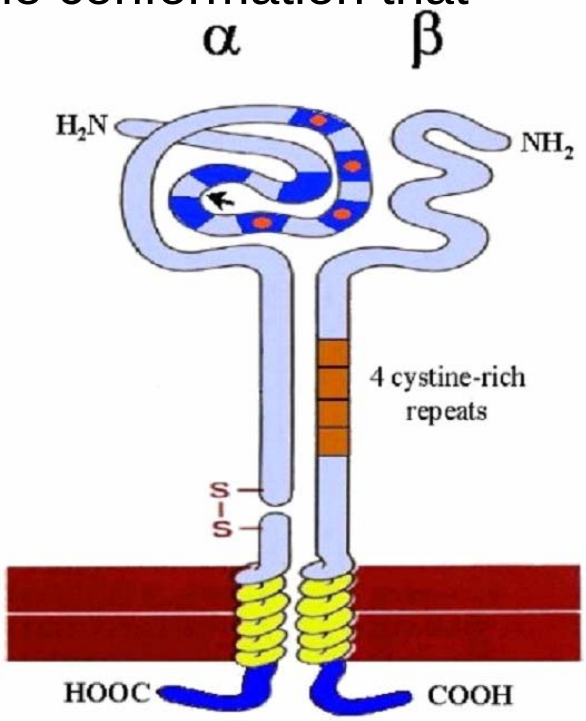


Biasing the conformation

Proteins are dynamic and constantly sample multiple conformations

Affinity for a target may be increased by stabilizing the conformation that is compatible with ligand binding

Integrin is a cell surface receptor that plays a role in cell-cell interaction and cell attachment to the extracellular matrix

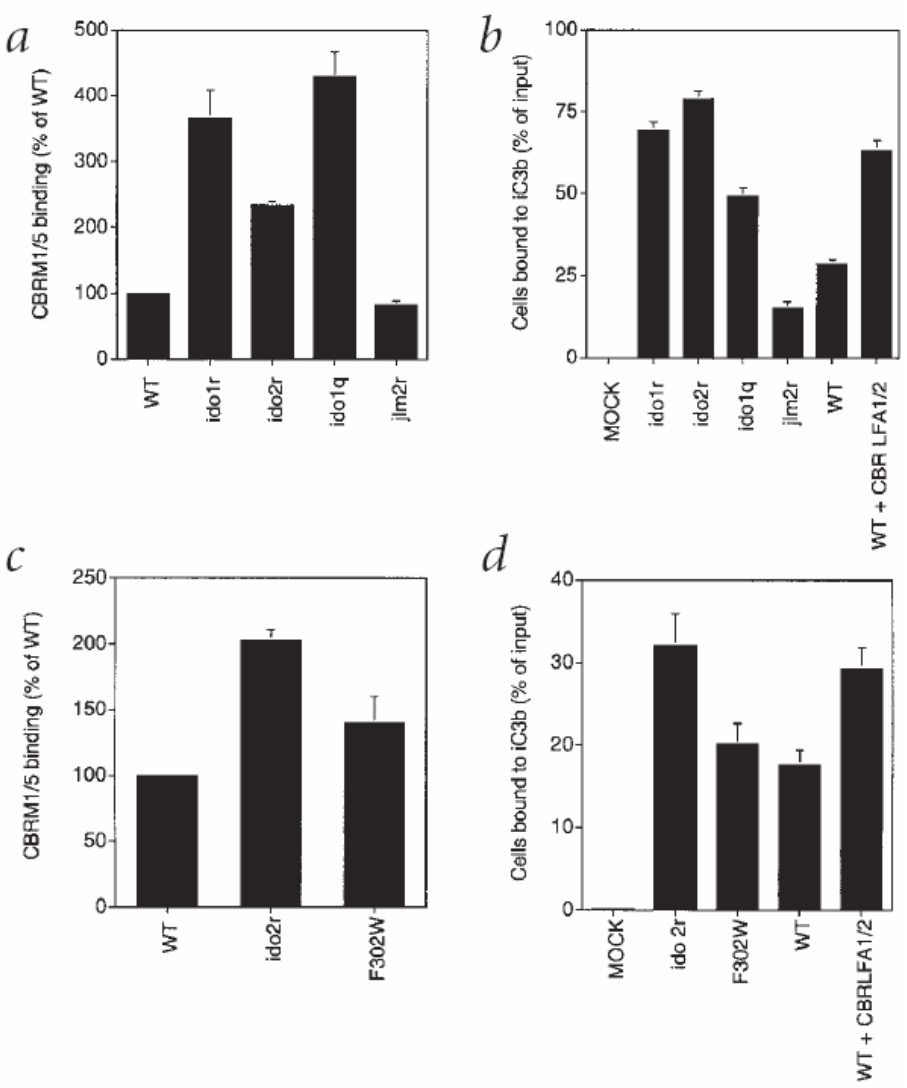


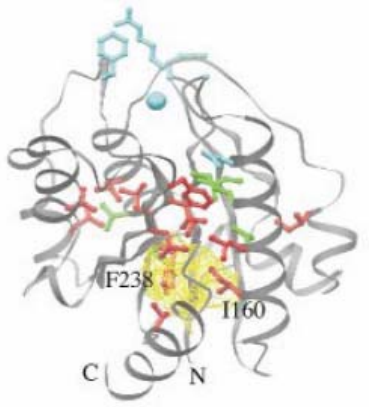
Binding to the ligand iC3b (a component of the C3 complement) is different between the **open** and closed conformations of the alpha subunit

Computationally stabilize either the open or closed conformation and test activity against the ligand

antibody binding

iC3b binding



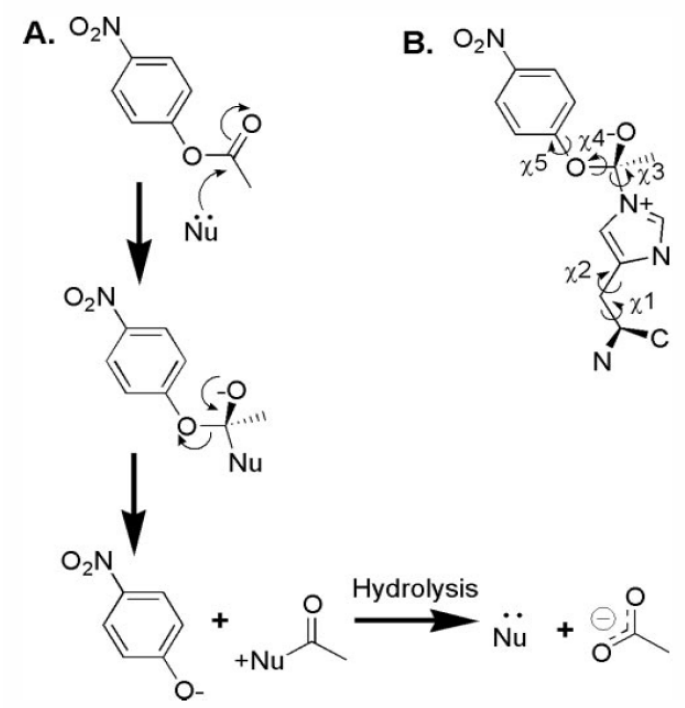


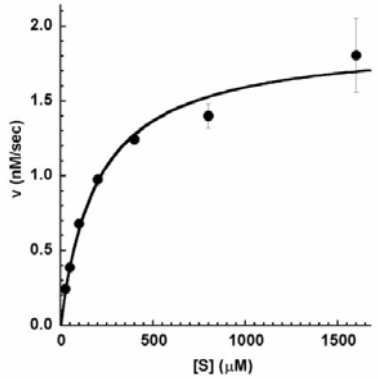
- four mutant sequences were computed using two different solvation potentials and subsets of core residues.
- mutations that stabilize the open conformation (1IDO) improves binding
- computationally designed mutant has a higher binding affinity than an “expert” designed mutant—F302W
- simply stabilizing a productive conformation can affect binding affinity

Shimaoka et al, NSB 7, 674 (2000)

Enzyme-like proteins

- Principles of enzymatic catalysis: proximity and orientation of substrate molecules, transition-state stabilization, acid base catalysis
- Model system
  - 108 residue protein rubredoxin mutant
  - histidine mediated hydrolysis of p-nitrophenylacetate (PNPA)
  - model the high energy transition state



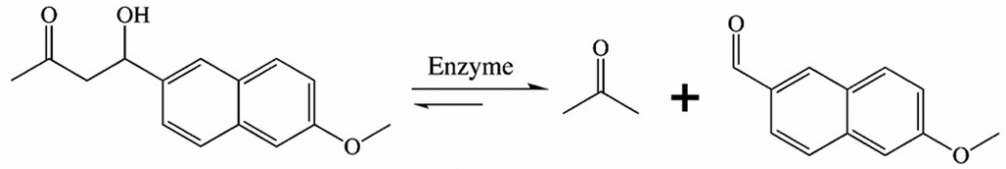


Design	Catalytic His position	Fraction hydrophobic exposure	Active site mutations
PZD1	12	0.11	F12H Y70A
PZD2	17	0.15	F12A L17H Y70A
PZD3	86	0.29	V86H I38A L42A L99A
PZD4	72	0.34	I72H L79A
PZD5	66	0.34	T66H F12A Y70A
PZD6	6	0.36	None
PZD7	39	0.37	A39H K57A
PZD8	91	0.39	V91H T77A
PZD9	49	0.39	Y49H K52A
PZD10	77	0.43	T77H L79A T89A

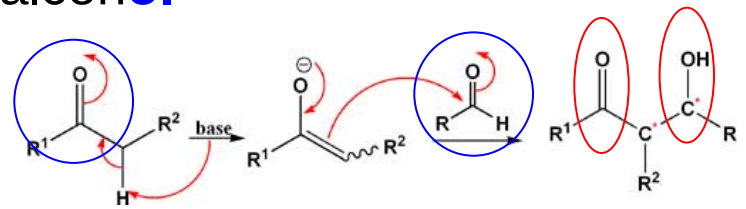
Bolon and Mayo, PNAS 98, 14274 (2001)

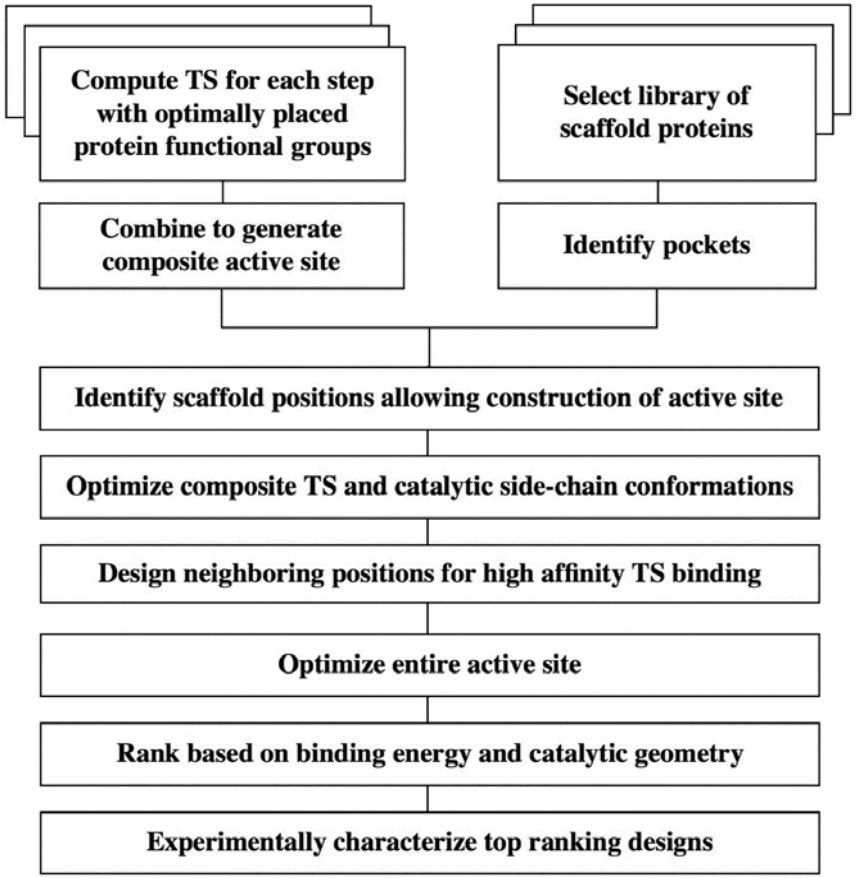
CPD of Retro Aldolase

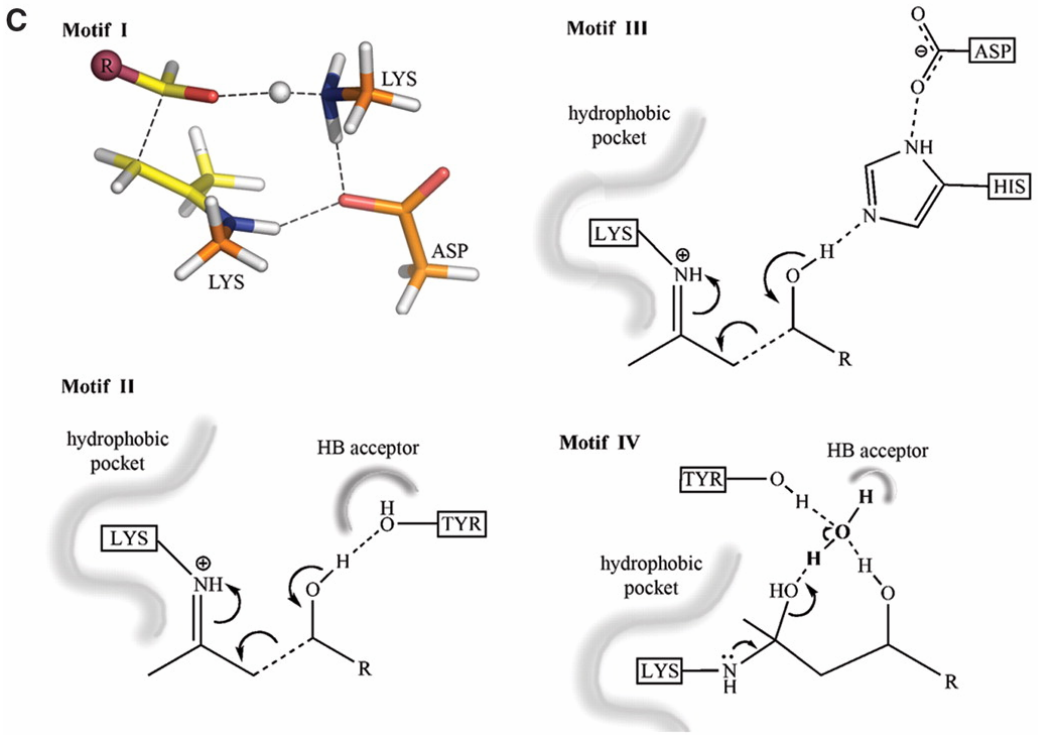
Retro aldol = reverse of aldol reaction



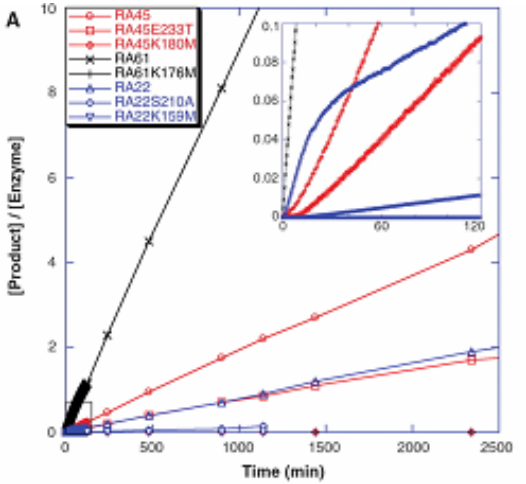
cf. aldol reaction = ketone (enolate) + **ald**ehyde ➔ketone + alcoh**ol**



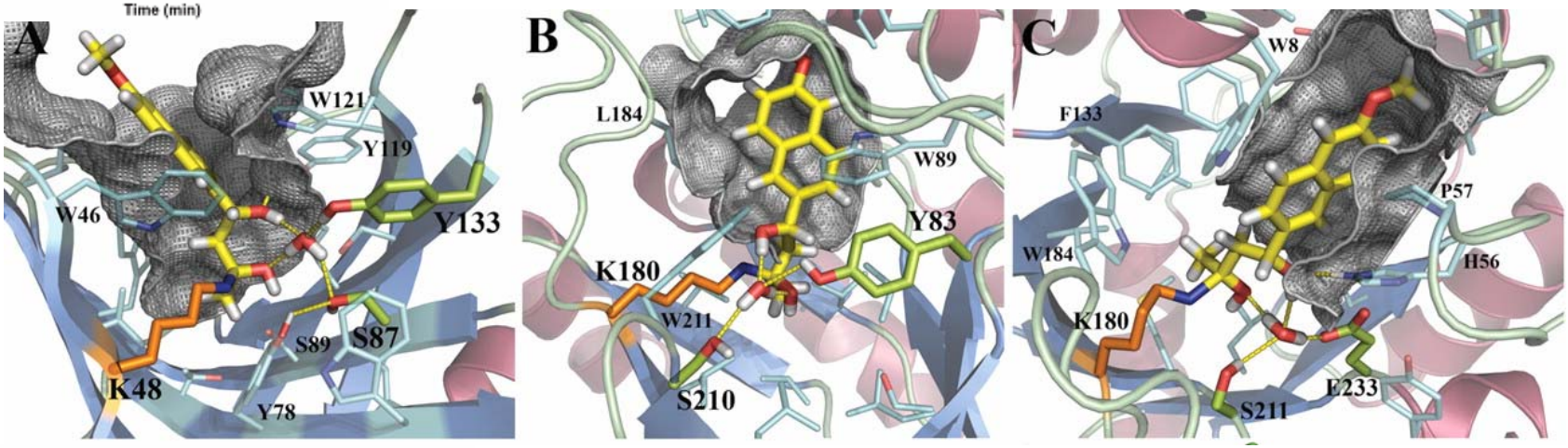


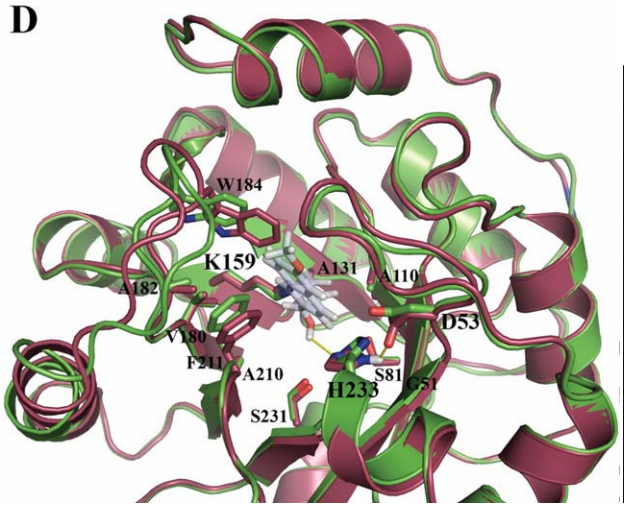


Jiang, et al. Science 319, 1387 (2008)



Motif	Catalytic lysine environment	Carbinolamine stabilization	Proton abstraction	Number tested	Number forming enaminone	Number of active designs	Rate enhancement
I	Polar	NC	Lys-Asp dyad	12	2	0	<4
II	Hydrophobic	NC	Tyr	9	1	0	<4
III	Hydrophobic	H-bond acceptor/donor	His-Asp dyad	13	10	10	10^2 10^3
IV	Hydrophobic	Water, H-bond acceptor	Water	38	20	22	10^3 10

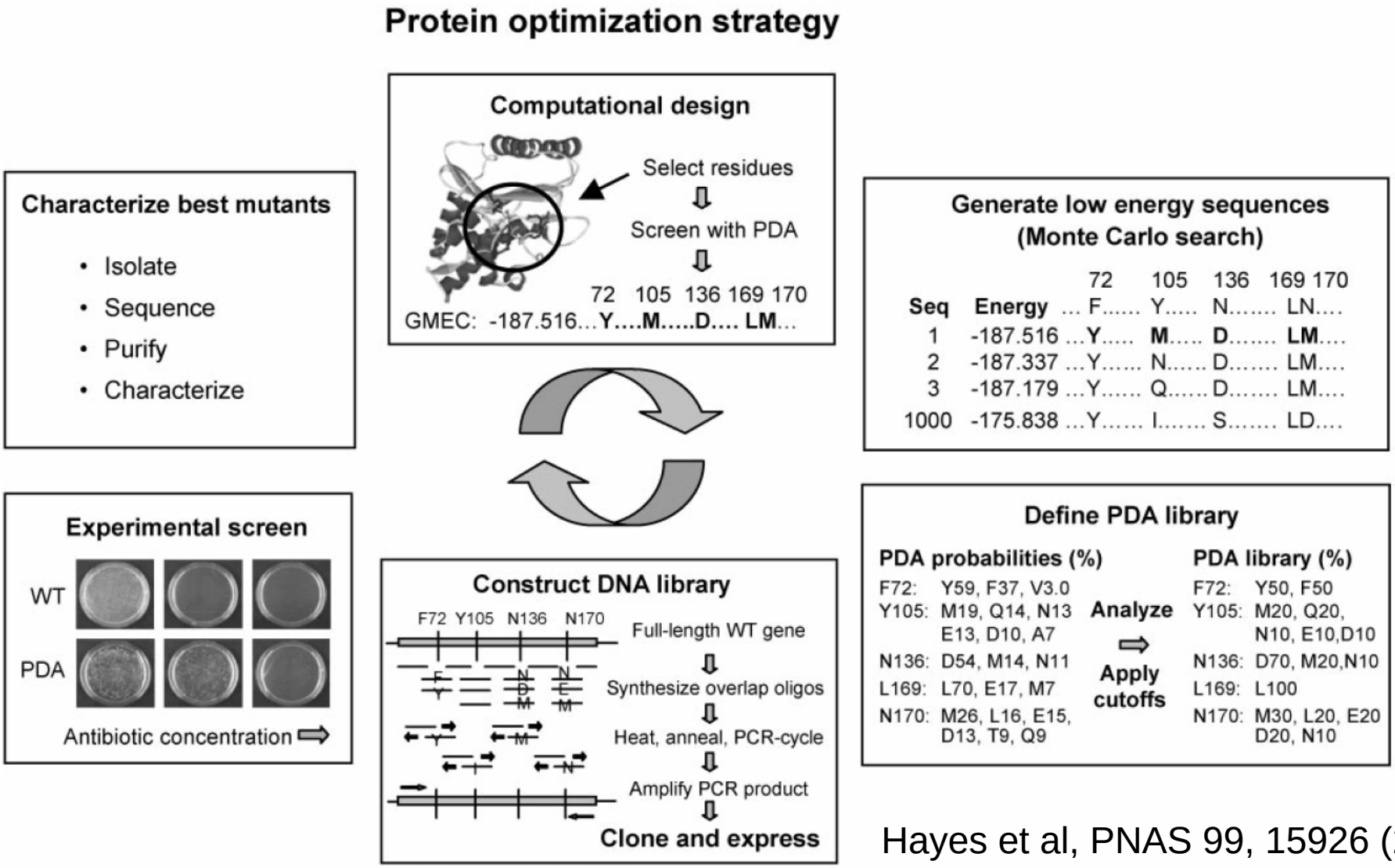




Design	k_{cat} (x 10^{-3} min^{-1})	K_M (μM)	k_{cat}/K_M (M^{-1} s^{-1})	k_{cat}/k_{uncat}*
RA22	3.1 ± 0.3 (b)	480 ± 130 (b)	0.11 ± 0.03 (b)	8.1 x 10^3 (b)
	0.5 ± 0.1 (s)	450 ± 210 (s)	0.018 ± 0.006 (s)	1.2 x 10^3 (s)
RA34	4.2 ± 1.1 (b)	620 ± 180 (b)	0.11 ± 0.01 (b)	1.1 x 10^4 (b)
	0.6 ± 0.1 (s)	600 ± 140 (s)	0.016 ± 0.004 (s)	1.5 x 10^3 (s)
RA45	2.3 ± 0.2	430 ± 48	0.091 ± 0.004	6.0 x 10^3
RA46	0.62 ± 0.5	290 ± 60	0.037 ± 0.006	1.6 x 10^3
RA60	9.3 ± 0.9	510 ± 33	0.30 ± 0.06	2.4 x 10^4
RA61	9.0 ± 1.0	210 ± 50	0.74 ± 0.11	2.3 x 10^4

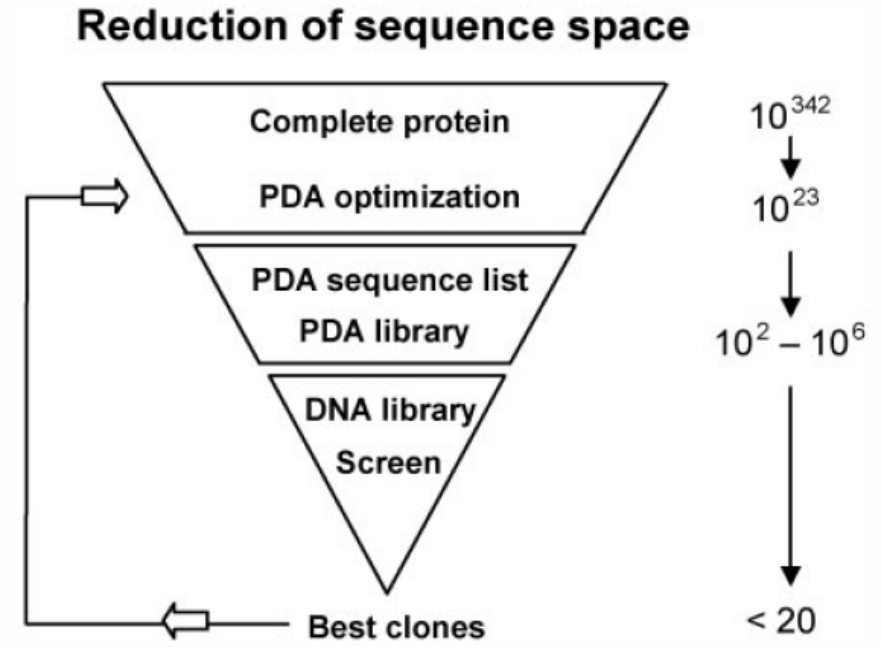
Computation-guided library

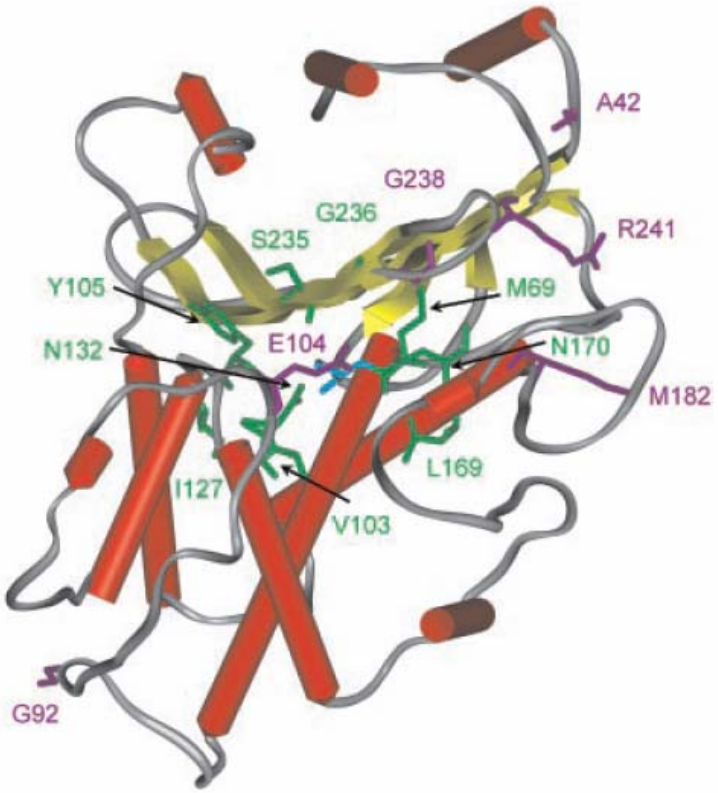
Computational predictions may be combined with a diversity oriented protein library to facilitate discovery



Hayes et al, PNAS 99, 15926 (2002)

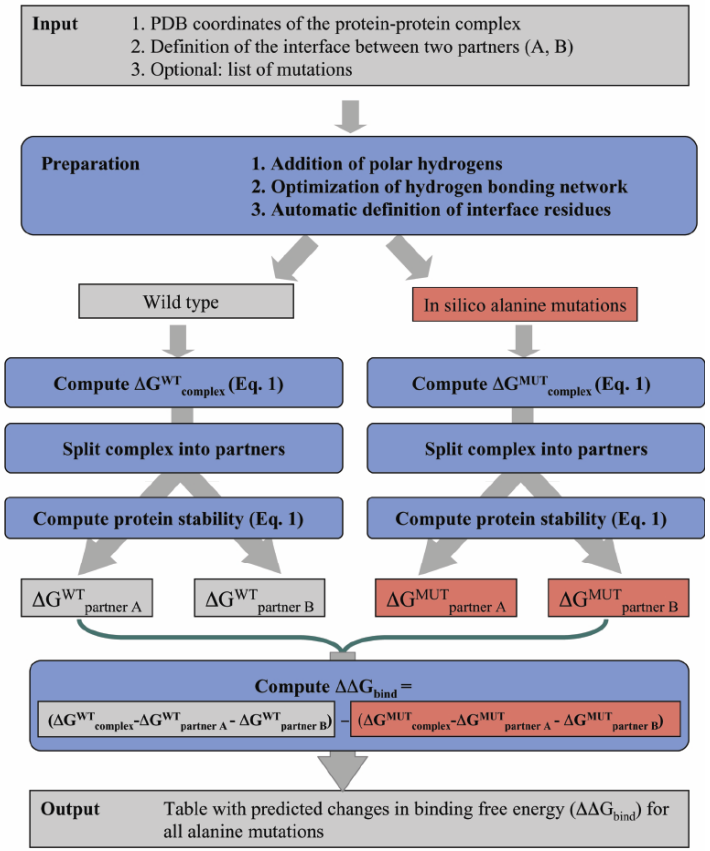
- Mutate 10 residues near the active site: 7 x 10^23 sequences
- Construct a library with the residue probabilities obtained from the 200,000 best mutants computed by DEE and MC
- Select for clones that survive on plates with high concentration of antibiotics
- 1,200 fold increase in resistance to antibiotic cefotaxime

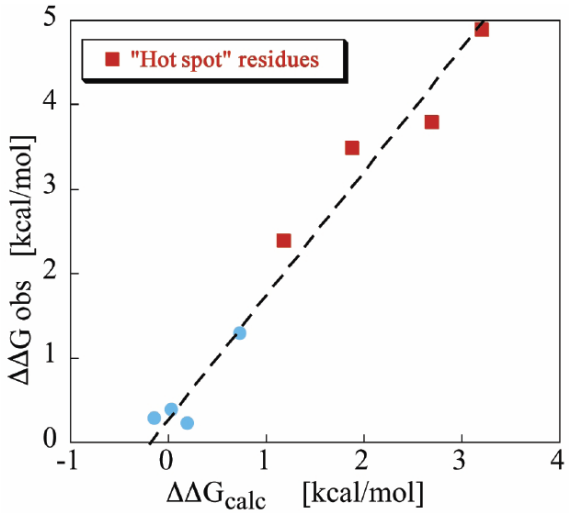


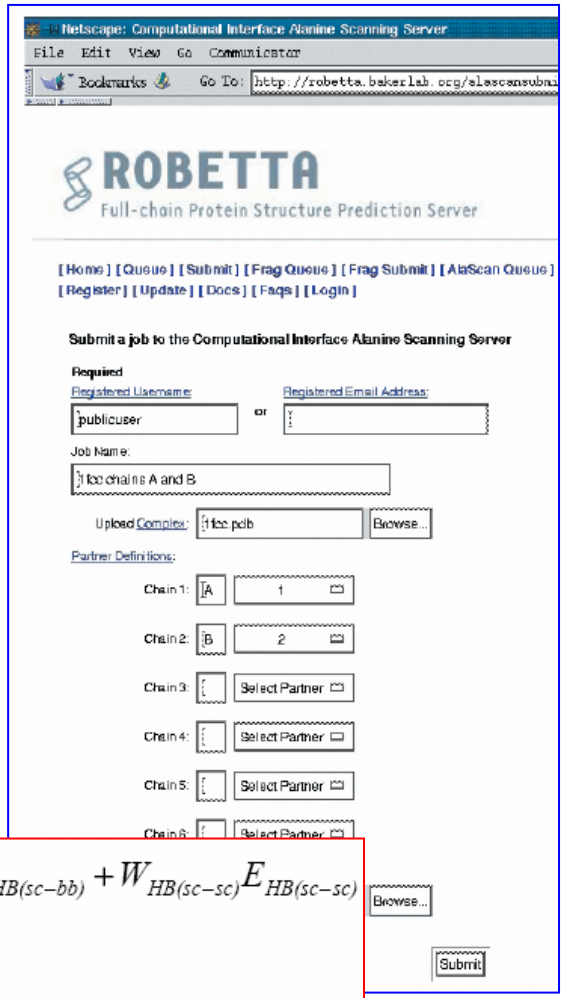


Computational alanine scanning

- Evaluate the thermodynamic consequence of making an ala mutation at the protein-protein interface by comparing the stability of complex with individual protein components
- Quickly identify “hot spots” comprising residues important for interaction







$$\Delta G = W_{attr}E_{LJattr} + W_{rep}E_{LJrep} + W_{HB(sc-bb)}E_{HB(sc-bb)} + W_{HB(sc-sc)}E_{HB(sc-sc)}$$
$$W_{sol}G_{sol} + W_{\phi/\psi}E_{\phi/\psi}(aa) + \sum_{aa=1}^{20} n_{aa}E_{aa}^{ref}$$

Kortemme et al, SciSTKE 219, l2 (2004)