DE NOVO DESIGN OF PROTEINS AND PROTEIN-PEPTIDE COMPLEXES: Advances and Challenges

Christodoulos A. Floudas

Stephen C. Macaleer '63 Professor in Engineering and Applied Science,
Professor of Chemical and Biological Engineering

Princeton University

Department of Chemical and Biological Engineering
Program of Applied and Computational Mathematics
Department of Operations Research and Financial Engineering
Center for Quantitative Biology

Outline

- **Motivational Examples**
- De Novo Protein Design: Definition
- De Novo Design: Background
  - Advances and Limitations
- Novel Two-Stage De Novo Protein Design Approach
  - Sequence Selection
    - Force Fields: Distance Dependent
    - Quadratic Assignment-like Models
    - Compstatin, Human beta defensin-2
  - Fold Specificity and Validation
    - First Principles based: Astro-Fold
    - NMR refinement like framework
- Computational Studies
  - Design of Inhibitors for Complement 3
  - Design of C3a
  - Design of Inhibitors for HIV-1, gp41, gp120
- Conclusions & Acknowledgements

Complement System

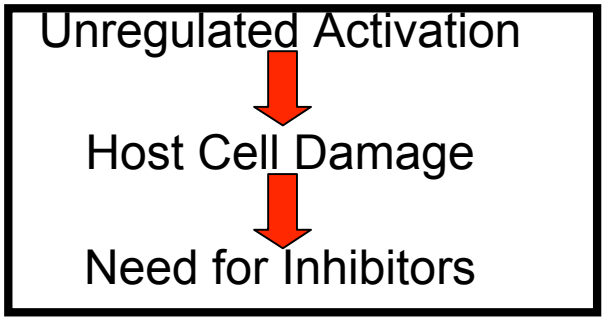
- ~30 distinct plasma proteins that interact to attack/eliminate pathogens
- Activated via (3) interacting pathways
  - (A) Classical : Antibody-binding (IgM, IgG) to pathogens
  - (B) Lectin : Mannose binding protein to carbohydrates on bacteria or viruses
  - (C) Alternative : Spontaneous binding to pathogens

    (A) : Adaptive/Acquired Immune Response

    (B); (C) : Innate/Natural Immune Response

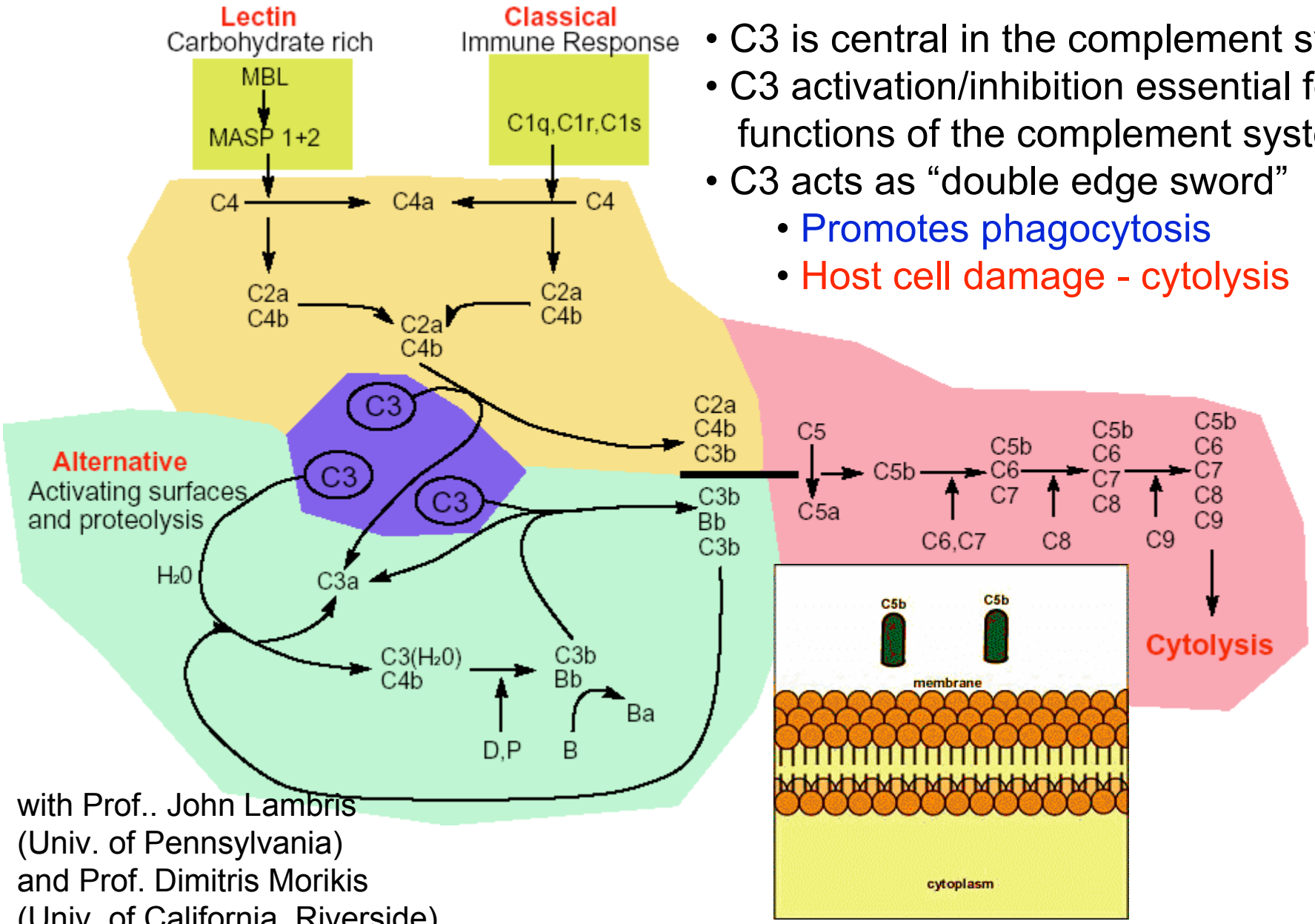
- Activation of the Complement System results in :
  - opsonization of pathogens (C3b; C4b)
  - recruitment of inflammatory cells (C3a; C5a; C4a+)
  - killing of pathogens (C5b, C6, C7, C8, C9)

    MAC: Membrane Attack Complex



• Acute Complement-Mediated Conditions	Annual US Patient Population
• Myocardial Infarction (Heart Attack)	1,500,000
• Coronary Artery Bypass	363,000
• Stroke	600,000
• Chronic Complement-Mediated Conditions	
• Rheumatoid Arthritis	2,100,000
• Alzheimer's Disease	4,000,000
• Systemic Lupus Erythematosus	500,000

Complement Pathways



- C3 is central in the complement system
- C3 activation/inhibition essential for all functions of the complement system
- C3 acts as "double edge sword"
  - Promotes phagocytosis
  - Host cell damage - cytolysis

with Prof.. John Lambris
(Univ. of Pennsylvania)
and Prof. Dimitris Morikis
(Univ. of California, Riverside)

Complement 3 : Design of Inhibitors

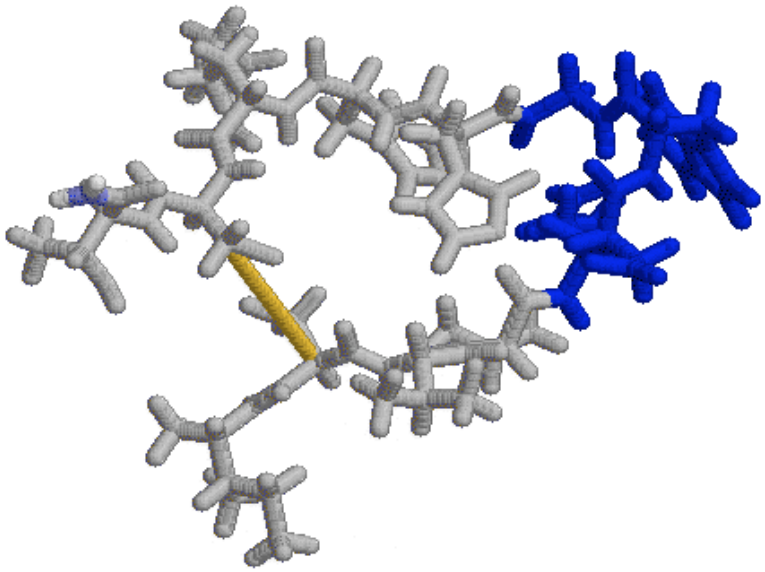
With Prof. John Lambris, University of Pennsylvania, School of Medicine
With Prof. Dimitris Morikis, University of California at Riverside

Compstatin : Synthetic Inhibitor

- 13 amino acid cyclic peptide
  ICVVQDWGHHRCT
- Disulfide bridge
- beta-turn

Objective

Designed improved Inhibitors
(Compstatin-like inhibitors)



C3a: Biologically active fragment of C3 component in the complement pathway

With Prof. Dimitris Morikis
(Univ. of California, Riverside)

A potent mediator of inflammation

Background

•77 residues, 3 S-S bonds, 4 α-helices

• C-terminal primary binding site (LGLAR)

• Super-potent peptide (12-15 times more active than natural C3a)
WWGKKYRASKLGLAR corresponding to positions 63-77 identified by Ember *et al*., Biochemistry, 1991

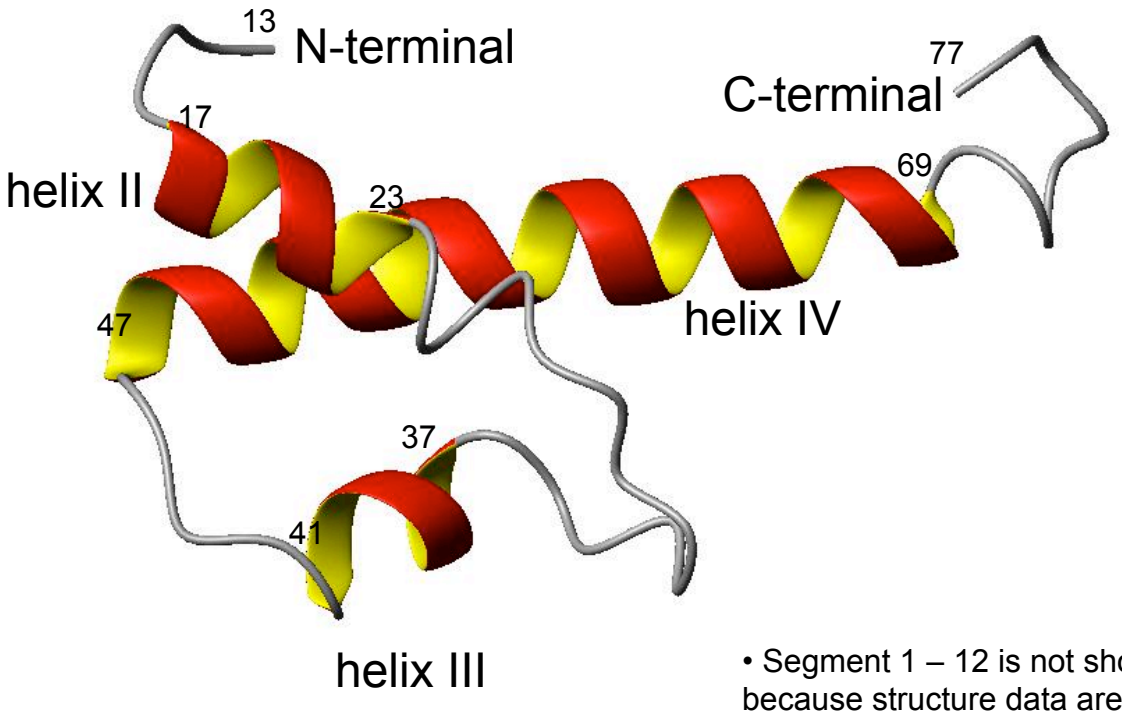
• Extensive sequence-activity studies by Ember *et al*., Biochemistry, 1991

• Ideal target for pharmaceutical development because of its small size and the fact that no complement inhibitor is yet available in clinic

Functions

Binds to C3a receptor (C3aR) with nanomolar affinity

Structure of C3a



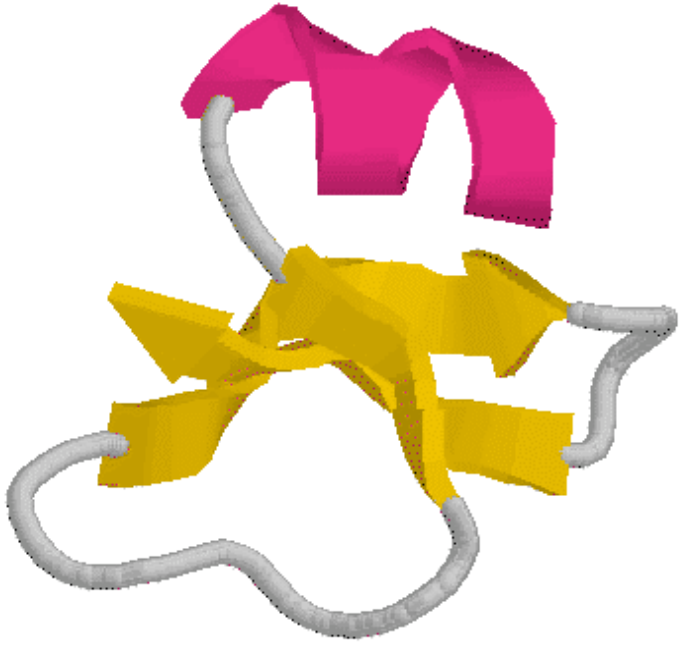
- Segment 1 – 12 is not shown because structure data are not available.
- helix I is segment 5 – 15.

Antibacterial Peptides

(with Prof. D. Morikis)

Beta-Defensins

- Family of antimicrobial peptides
- Cationic peptides of 28-42 AAs
- Structure for only (2) humanBDs
  - Structure-function unknown
  - Low sequence identity
- hBD-2 10x more potent than hBD-1



	25					30				35				40				44
hBD-2	I	G	D	P	V	T	C	L	K	S	G	A	I	C	H	P	V	F	C	P
hBD-1	.	.	D	H	Y	N	C	V	S	S	G	G	Q	C	L	Y	S	A	C	P
mBD-7	.	N	S	K	R	A	C	Y	R	E	G	G	E	C	L	Q	.	R	C	I
mBD-8	.	N	E	P	V	S	C	I	R	N	G	G	I	C	Q	Y	.	R	C	I
hBD-3	T	L	Q	K	Y	Y	C	R	V	R	G	G	R	C	A	V	L	S	C	L
hBD-4	F	E	L	D	R	I	C	G	Y	G	T	A	R	C	R	K	.	K	C	R
mBD-1	.	.	D	Q	Y	K	C	L	Q	H	G	G	F	C	L	R	S	S	C	P
mBD-2	.	A	E	L	D	H	C	H	T	N	G	G	Y	C	V	R	A	I	C	P
mBD-3	I	N	N	P	V	S	C	L	R	K	G	G	R	C	W	N	.	R	C	I
mBD-4	I	N	N	P	I	T	C	M	T	N	G	A	I	C	W	G	.	P	C	P
bBD-1	.	.	D	F	A	S	C	H	T	N	G	G	I	C	L	P	N	R	C	P
bBD-2	.	.	N	H	V	T	C	R	I	N	R	G	F	C	V	P	I	R	C	P
bBD-12	.	.	G	P	L	S	C	G	R	N	G	G	V	C	I	P	I	R	C	P

	45					50				55				60				64
hBD-2	R	R	Y	K	Q	I	G	T	C	G	L	P	G	T	K	C	C	K	K	P
hBD-1	I	F	T	K	I	Q	G	T	C	Y	R	G	K	A	K	C	C	K	.	.
mBD-7	G	L	F	H	K	I	G	T	C	.	N	F	R	F	K	C	C	K	F	Q
mBD-8	G	L	R	H	K	I	G	T	C	.	G	S	P	F	K	C	C	K	.	.
hBD-3	P	K	E	E	Q	I	G	K	C	S	T	R	G	R	K	C	C	R	R	K
hBD-4	S	Q	E	Y	R	I	G	R	C	P	N	T	Y	A	.	C	C	L	R	K
mBD-1	S	N	T	K	L	Q	G	T	C	K	P	D	K	P	N	C	C	K	S	.
mBD-2	P	S	A	R	R	P	G	S	C	F	P	E	K	N	P	C	C	K	Y	M
mBD-3	G	N	T	R	Q	I	G	S	C	G	V	P	F	L	K	C	C	K	R	K
mBD-4	T	A	F	R	Q	I	G	N	C	G	H	F	K	V	R	C	C	K	I	R
bBD-1	G	H	M	I	Q	I	G	I	C	F	R	P	R	V	K	C	C	R	S	W
bBD-2	G	R	T	R	Q	I	G	T	C	F	G	P	R	I	K	C	C	R	S	W

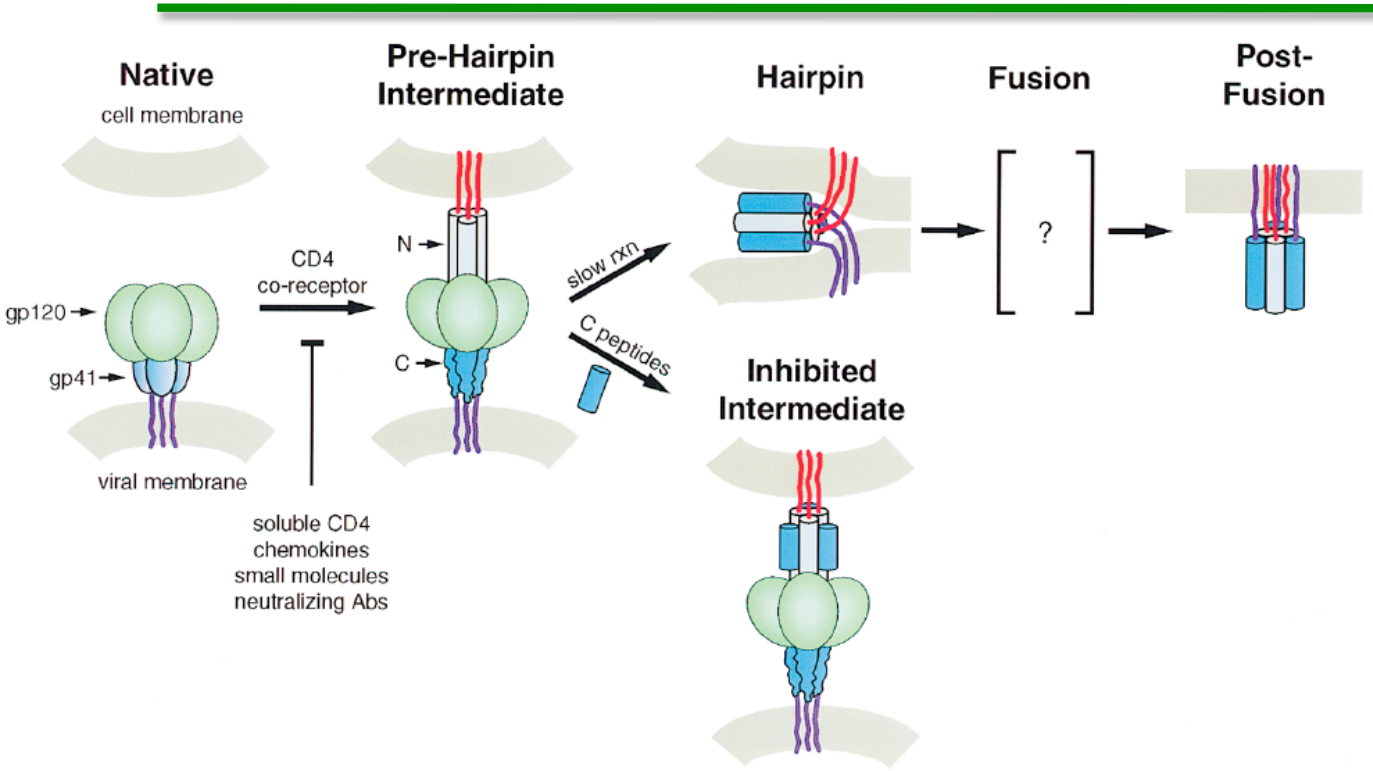
Objective

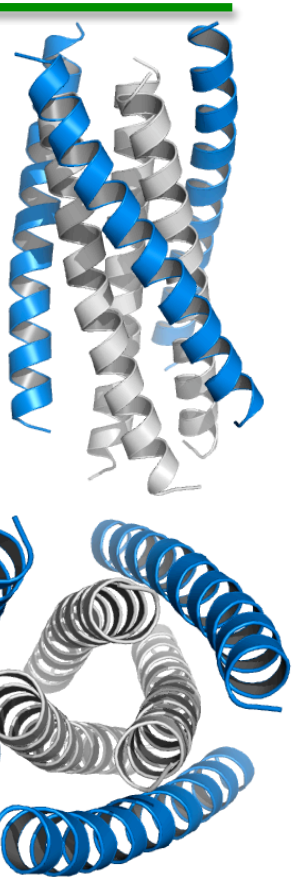
Design improved antibacterial peptides

De Novo Design of anti-HIV-1 entry peptides

with Professor **Robert Siliciano**
Johns Hopkins School of Medicine

HIV Membrane Fusion





N-Helix: NH2-MTLTVQSRQLLSGIVQQQNNLLRAIEAQQHLLQLTVWGIKQLQARILAVERYLKDQ-Ac
535 590

C-Helix: NH2-**WMEWDREINNYTSLIHSLIEESQNQQEKNEQELLELDKWASLWNWF**-Ac
682 673

T-20: **YTSLIHSLIEESQNQQEKNEQELLELDKWASLWNWF** (36 amino acids)

C34: **WMEWDREINNYTSLIHSLIEESQNQQEKNEQELL**

Design of HIV-1 gp41 inhibitor

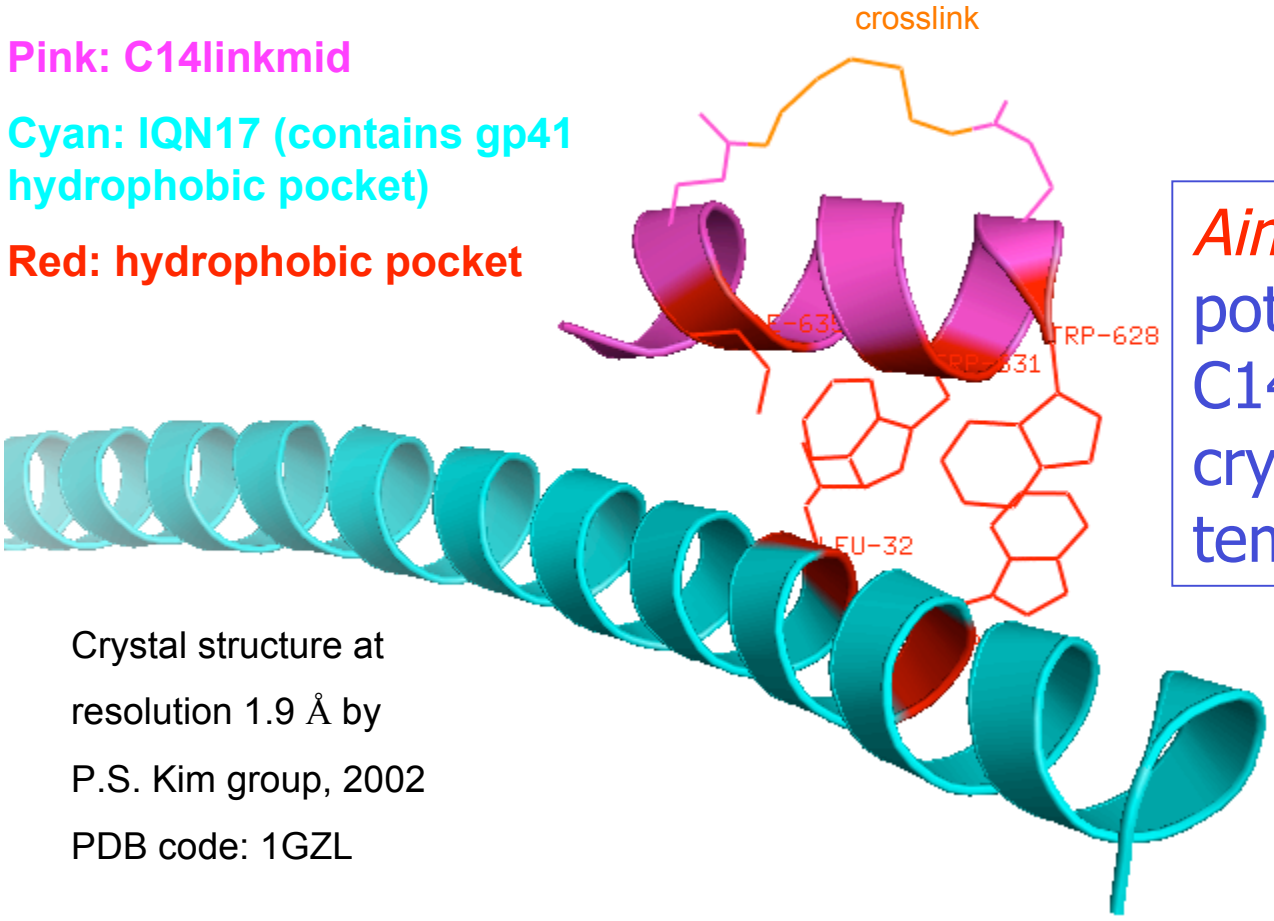
Background on gp 41

•Crystal structure of C14linkmid bound to the gp41 hydrophobic pocket

Pink: C14linkmid

Cyan: IQN17 (contains gp41 hydrophobic pocket)

Red: hydrophobic pocket



Crystal structure at resolution 1.9 Å by P.S. Kim group, 2002

PDB code: 1GZL

Aim: design a more potent inhibitor than C14linkmid using this crystal structure as template

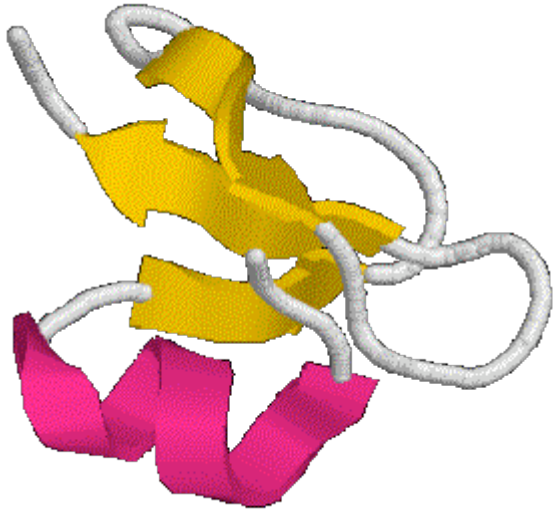
Outline

- Motivational Examples
- **De Novo Protein Design: Definition**
- De Novo Design: Background
  - Advances and Limitations
- Novel Two-Stage De Novo Protein Design Approach
  - Sequence Selection
    - Force Fields: Distance Dependent
    - Quadratic Assignment-like Models
    - Compstatin, Human beta defensin-2
  - Fold Specificity and Validation
    - First Principles based: Astro-Fold
    - NMR refinement like framework
- Computational Studies
  - Design of Inhibitors for Complement 3
  - Design of C3a
  - Design of Inhibitors for HIV-1, gp41, gp120
- Conclusions & Acknowledgements

De Novo Protein Design

Define target template

Backbone coordinates for N,Ca,C,O and possibly Ca-Cb vectors from PDB

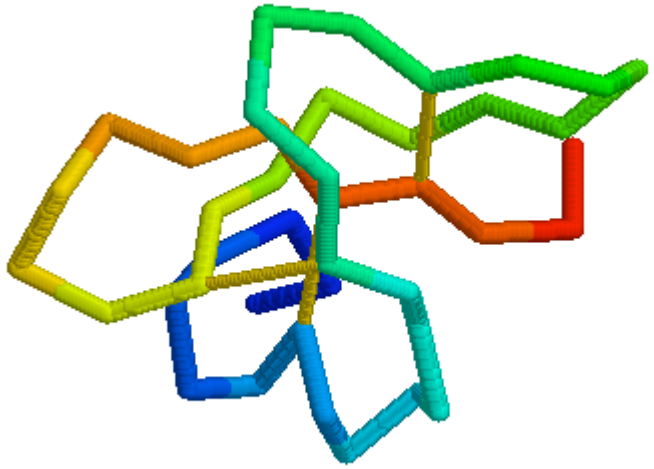


Human β-Defensin-2

hbd-2 (PDB: 1fqq)

Design folded protein

Which amino acid sequences will stabilize this target structure ?



Full sequence design

Mayo et al.; Hellinga et al.; DeGrado et al; Saven et al.; Hecht et al.

Combinatorial complexity

-Backbone length : n

-Amino acids per position : m

m^n possible sequences

Challenges

In silico sequence selection

Fold validation/specificity

De Novo Protein Design: challenges

- **Flexibility of Backbone Templates**
- **Full sequence combinatorial design of proteins of practical size still challenging**

Full combinatorial design of a 100-residue protein

20^{100} or 10^{130} amino acid sequences,

$(20r)^{100}$ rotamer sequences to consider !

Average number of rotamers per amino acid

- Currently often only possible to design core, boundary or surface regions of small protein domains (25 – 74 residues) (Gordon *et al.*, J. Comput. Chem., 2003)

- **De novo protein design: NP-hard problem**

- No exact polynomial-time algorithms known

- For exponential-time algorithms, computation time varies exponentially with number of design positions

Pierce and Winfree, 2002
Fung, Rao, Floudas, Prokopyev, Pardalos, Rendl, 2005

Outline

- Motivational Examples
- De Novo Protein Design: Definition
- **De Novo Design: Background**
  - **Advances and Limitations**
- Novel Two-Stage De Novo Protein Design Approach
  - Sequence Selection
    - Force Fields: Distance Dependent
    - Quadratic Assignment-like Models
    - Compstatin, Human beta defensin-2
  - Fold Specificity and Validation
    - First Principles based: Astro-Fold
    - NMR refinement like framework
- Computational Studies
  - Design of Inhibitors for Complement 3
  - Design of C3a
  - Design of Inhibitors for HIV-1, gp41, gp120
- Conclusions & Acknowledgements

Background and Advances

- Stochastic Methods: MC, Genetic Algorithms

  Tuffery et al. (1991); Desjarlais, Handel (1995)(2003)

- Probabilistic Approaches, Combinatorial Libraries

  Saven & coworkers (2000)(2001)(2004)

- Deterministic Methods

  - Self-Consistent Mean Field (Koehl, Delarue, 1994)

  - Self-Consistent Mean Field and MC (Koehl, Levitt, 1999a.b)

  - Dead End Elimination Criteria (Mayo & coworkers; Desjarlais, Handel, 1995, 1999; Hellinga & coworkers; Desmet et al. 1992; Goldstein, 1994; Pierce et al. 2000; Gordon et al. 2003; Gregoriev et al, 2006; Gregoriev, Donald, 2007)

- Iterative Sequence-Structure (Kuhlman et al. 2003; Saunders & Baker, 2005)

- Quadratic Assignment-like models for sequence selection and first principles fold validation (Klepeis et al., 2003, 2004; Fung et al. 2005, 2007)

Advances: Towards Flexible Backbone Templates_

- **Scaling down the atomic van der Waals radii by a factor (~5-10%)**
  - **Overestimation of attractive forces between atoms and the possibility of atom overpacking**

Dahiyat, B.I. & Mayo, S.L. (1997) PNAS

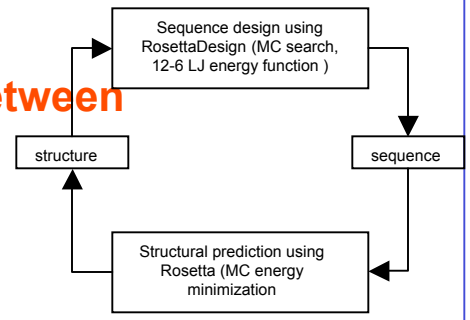
- **Considering a fixed set of rotamers (DEE) or changing super-secondary structure parameters which alter relative orientation and distance between secondary structures**
  - **Only a subset of possible conformations is considered**

Su, A. & Mayo, S.L. (1997) Protein Sci.

- **Generating ensemble of random structures from template**
  - Solve each structure in the ensemble assuming fixed backbone and apply genetic algorithms and Monte Carlo sampling to combine results into a single low energy structure
  - **Only a random subset of possible conformations is considered**

Desjarlais & Handel (1999) *J. Mol. Bio.*

- **Iterating between sequence space and structure space**
  - **Backbone flexibility only indirectly addressed by transitions between similar structures in the structure space**



Saunders & Baker (2005) *J. Mol. Bio.*

Kuhlman, Dantas, Ireton, Varani, Stoddard, & Baker (2003)*J. Mol. Bio.*

- **Dead End Elimination with Flexible Templates**

Gregoriev and Donald, 2007, Bioinformatics

Flexibility of Backbone Templates

De novo protein design framework allows true backbone flexibility

$$d^{L}_{C^{\alpha}-C^{\alpha}} \le d_{C^{\alpha}-C^{\alpha}} \le d^{U}_{C^{\alpha}-C^{\alpha}}$$

$$\phi^{L} \le \phi \le \phi^{U}$$

$$\varphi^{L} \le \varphi \le \varphi^{U}$$

Incorporated in stage 1 through the use of distance bins and in stage 2 through lower and upper bounds.

Incorporated in stage 2 through the template-constrained folding calculation. Bounds are ± 10°.

- **C^{α}-C^{α} distance and dihedral angles are bounded continuous functions**

Floudas, AIChE J. (2005); Klepeis, Floudas, Morikis, Lambris, JACS (2003), IERC (2004); Fung, Rao, Floudas, Prokopyev, Pardalos, Rendl, J. Comb. Optim. (2005)Fung, Taylor, Floudas, Opt. Meth. Soft. (2007)

- **NMR ensemble**
- **MD with GB**
- **MD with Explicit water molecules**

Outline

- Motivational Examples
- De Novo Protein Design: Definition
- De Novo Design: Background
  - Advances and Limitations
- **Novel Two-Stage De Novo Protein Design Approach**
  - **Sequence Selection**
    - **Force Fields: Distance Dependent**
    - **Quadratic Assignment-like Models**
    - **Compstatin, Human beta defensin-2**
  - **Fold Specificity and Validation**
    - **First Principles based: Astro-Fold**
    - **NMR refinement like framework**
- Computational Studies
  - Design of Inhibitors for Complement 3
  - Design of C3a
  - Design of Inhibitors for HIV-1, gp41, gp120
- Conclusions & Acknowledgements

De Novo Protein Design Framework

Sequence selection stage: generates a rank-ordered list of sequences with the lowest energies by solving an integer linear programming (ILP) model

- Quadratic Assignment-like models

(Klepeis et al., *JACS* (2003); Klepeis et al., *IERC* (2004); Fung *et al*., *J. Comb. Optim*.,2005; Fung et al., *OMS*, 2007; Fung et al., *Biophysical J*., 2008)

- Distance-dependent C^{α}-C^{α}, centroid-centroid forcefields

(Loose, Klepeis, Floudas, *Proteins* (2004); Rajgaria, McAllister, Floudas, *Proteins*, 2006; Rajgaria et al., *Proteins,* 2008)

Fold specificity stage: calculates specificity of each sequence to the flexible design template using full-atomistic force fields AMBER, ECEPP/3

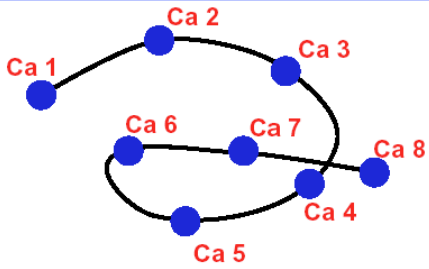
- First principles via ASTRO-FOLD

(Klepeis and Floudas, *Biophys. J*., 2003)

- NMR structures refinement-based method via CYANA and TINKER using AMBER (Fung et al, *Biophys. J.,* 2008; Taylor et al. *Biophys. J.,* 2008)

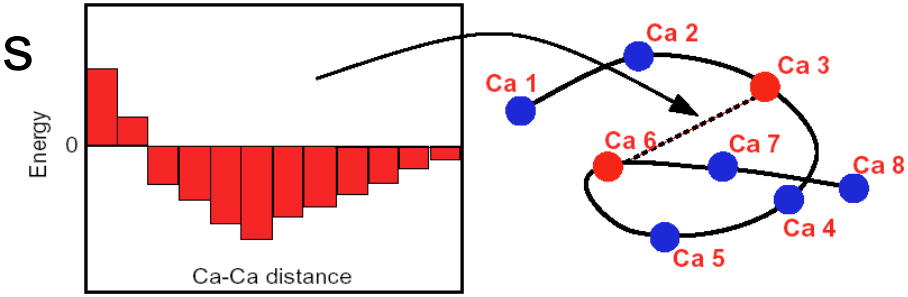
De Novo Protein Design Framework

Klepeis, Floudas, Lambris, Morikis 2003, 2004
Fung, Taylor, Floudas 2005, 2007, 2008



Sequence selection

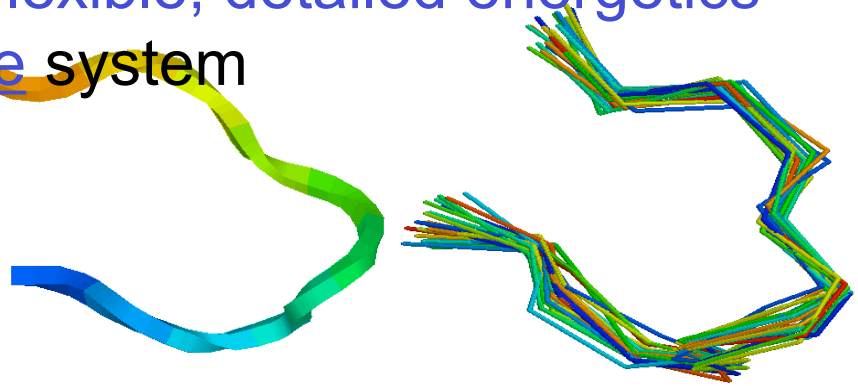
- Identify target template for desired fold; specify coordinates of backbone
- Identify possible residue mutations
- ***Distance dependent pairwise potentials***
- Generate rank-ordered energetic list from mixed-integer linear (MILP)



1	2	3	4	5	6	7	8
A	T	R	E	G	F	A	Q
A	S	K	E	P	Y	G	Q
V	S	K	E	G	F	A	Q

Fold Validation:Specificity

- Model selected sequences using flexible, detailed energetics
- Employ global optimization for free system
- Employ global optimization for system constrained to template
- Calculate relative probability for structures similar to desired fold



A High Resolution Ca-Ca and Side Chain Centroid Based Distance Dependent Force Field

R. Rajgaria S. R. McAllister and C. A. Floudas, *Proteins, 70, 950-970 (2008)*
R. Rajgaria S. R. McAllister and C. A. Floudas, *Proteins, 65(2), 726-741 (2006)*
C. Loose, J.L. Klepeis and C.A. Floudas, *Proteins, 54:303-314 (2004)*

Objectives

- Create a distance-dependent force field to find native protein folds.
- Design a training procedure that will make the force field robust using large scale linear optimization
- Test our force field against a very good distance dependent force fields (e.g., TE-13)[1] by attempting to identify the native fold of novel proteins.

1 Tobi, D.; Elber, R. Distance-Dependent, Pair Potential for Protein Folding. *Proteins: Structure, Function, and Genetics* **2000**, 41, 40-46.

Force Field – Formulation*

$C^{\alpha}-C^{\alpha}$ distance dependent

8-bin definition (ID)

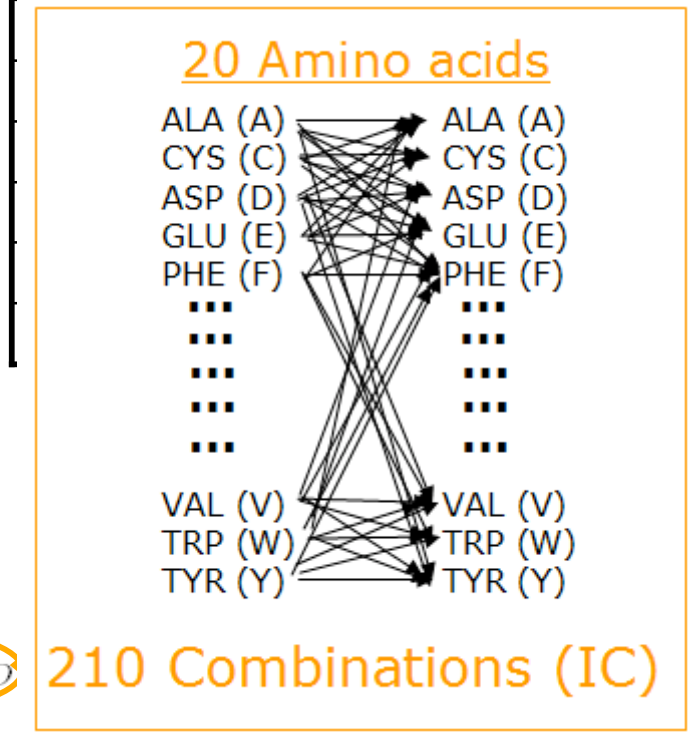
More resolution for bin 3 to 6

210 amino-acid combination (IC)

1680 energy variables $\theta_{IC,ID}$

Table 1: Bin Definition

Bin ID	C^{α}-distance [A°]
1	3-4
2	4-5



Energy calculation

Sum of pairwise interaction at a particular distance

$$E(X_{p,i}) = \sum_{IC} \sum_{ID} N_{p,i,IC,ID}$$

Parameter

*Loose. C., Klepeis. J.L., and Floudas. C.A., Proteins ,2004, 54, 303-314.

*Rajgaria. R., McAllister. S., and Floudas C.A. Proteins,2006, 65, 726-741 .

Force Field – Formulation*

Anfinsen's hypothesis was used as main criteria for energy evaluation

$$E(X_{p,i}) - E(X_{p,n}) > \varepsilon \quad p = 1, \ldots, P \quad i = 1, \ldots, N$$

$$\sum_{IC} \sum_{ID} [N_{p,i,IC,ID} - N_{p,n,IC,ID}]\theta_{IC,ID} + S_p \geq \varepsilon$$

$$p = 1, \ldots, P \quad i = 1, \ldots, N$$

$$\theta_{IC,\,ID} \in [-25,25]$$

$$\min_{\theta(IC,ID)} \sum_{p} S_p$$

Many more constraints – based on physical properties of interacting amino acids

*Loose. C., Klepeis. J.L., and Floudas. C.A., Proteins ,2004, 54, 303-314.

*Rajgaria. R., McAllister. S., and Floudas C.A. Proteins, 2006, 65, 726-741.

Constraints

Smooth profile (Tobi and Elber, 2000)

$$\theta_{IC,ID+1} - \theta_{IC,ID} \geq -8, \quad \forall IC; ID = 1$$

Decrease in effectiveness at long distances

$$\theta_{IC,ID} \leq 5, \quad \forall IC; ID = 7$$

Favorable interaction at 4-6.5 Å between hydrophobic groups
(Bahar and Jernigan, 1997)

$$\theta_{IC,ID} \leq 0, \quad IC \in \{H,H\}; ID = 2,3,4,5$$

"Energy well" formation at around 4.5 to 5.0 Å (below this "steric effects" and above this "insufficient contact")

$$\theta_{IC,ID+1} - \theta_{IC,ID} \leq -4, \quad IC \in \{H,H\}; ID = 1$$

$$\theta_{IC,ID+1} - \theta_{IC,ID} \leq -2, \quad IC \in \{H,H\}; ID = 2,3$$

High Resolution Decoys - Generation

The Set: 1482 non-homologous proteins from Skolnick and co-workers*.

Method

- Identify hydrophobic residues in the secondary structure

– If protein contains little secondary structure then consider all hydrophobic residues.

– Introduce a range of distance variations for the selected residues (8 values between 0.5 Å and 5.0 Å).

– An ensemble of 200 structures is created through torsion angle dynamics enforcing the distance bounds on the hydrophobic core.

*Zhang. Y. and Skolnick J. PNAS, 2004, 101, 7594-99.

Method and Implementation

Training

1250 proteins and 500 decoys of each protein

LP formulation to optimize energy parameters

Due to limited computer memory

Only a small subset of high quality decoys were used at a time

Iterative dropping scheme was used to include all decoys in force field generation

Testing

High Resolution Set

150 randomly selected proteins

500 decoys of each protein

RMSD (native)	Training Set	Test Set
0.0-0.5	**12**	**1**
0.5-1.0	**458**	**60**
1.0-1.5	**607**	**74**
1.5-2.0	**173**	**15**

Medium Resolution Set

151 randomly selected proteins

200 decoys of each protein

Minimum RMSD distribution

RMSD (native)	Test Set
3.0-16.0	**150**

Results* – Testing the Force Field

Evaluation Metrics

- average rank
- number of first ranked proteins
- average RMSD
- Z-score

$$Z = \frac{< E > - E_n}{\sqrt{< E^2 > - < E >^2}}$$

Test on High Resolution Decoys

FF name	Avg. Rank	# Firsts	Avg. RMSD (Å)	Avg. Z-score
HR	1.87	113	0.45	2.11
LKF	39.45	17	1.72	1.55
TE-13	19.94	92	0.81	3.15

**Rajgaria. R., McAllister. S., and Floudas C.A. Proteins, 2006, 65, 726-741.*

Results – Testing the Force Field

Test on Medium Resolution Decoys

Common proteins between HR training set and
LKF test set were removed

FF name	Avg. Rank	# Firsts	Avg. Z-score	Avg.RMSD(Å)
HR	4.32	86/110*	3.83	1.90
LKF	5.84	93/151	3.08	3.51
TE-13**	17.36	43/131	2.01	- not avail -

Tobi. D., and Elber. R., Proteins ,2000, 41, 40-46.

Side Chain Based Force Field

Importance

– C^{α} based formulation disregards presence of the side chain atoms

– Inclusion of side chain atoms might improve the energy estimation

– Side chain dependence needed for protein design problems

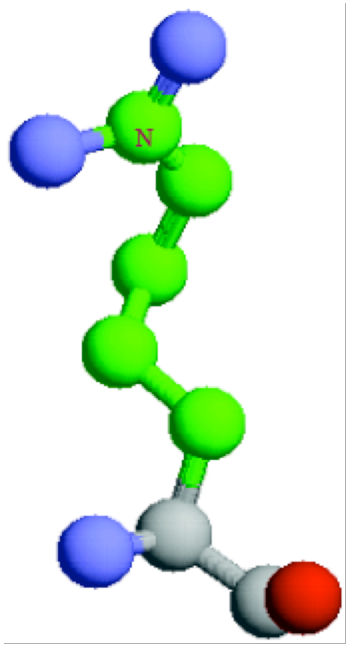
Need to revisit the interaction center definition

"effective" distance range might be different

define "centroid"

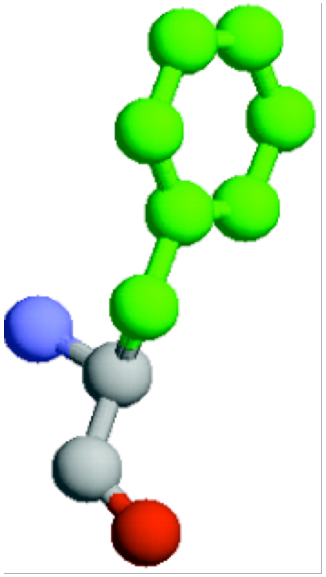
Force Field – Side Chain Based Formulation

Side Chain Centroid definition

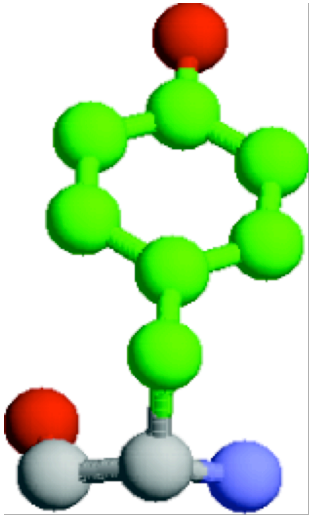


Arg

Pro



Phe



Tyr

Force Field – Side Chain Based Formulation

Side Chain Centroid distance dependence

6-bin definition (ID)

More resolution for bin 2 to 5

210 amino-acid combination (IC)

1260 energy variables ($\theta_{IC,ID}$

Energy calculation

Sum of pairwise interaction

$$E(X_{p,i}) = \sum_{IC} \sum_{ID} N_{p,i,IC,ID}\, \theta_{IC,ID}$$

Table 5: 6-Bin Definition

Bin ID	Centroid-distance[Aº]
1	4-5
2	5-5.5
3	5.5-6
4	6-6.5
5	6.5-7
6	7-8

Table 6: 7-Bin Definition

Bin ID	Centroid-distance[Aº]
1	4-5
2	5-5.5
3	5.5-6
4	6-6.5
5	6.5-7
6	7-8
7	8-9

Results – Testing the Force Field

Testing Centroid Based Force Field on High Resolution Decoys

FF name	Avg. Rank	# Firsts	Avg.RMSD (Å)	Avg. Z-score
6bin-HRSC	2.49	128/148	0.29	3.62
7bin-HRSC	2.01	125/148	0.32	3.39
HR	1.87	113/150	0.45	2.11
LKF	39.45	17/150	1.72	1.55
TE-13*	19.94	92/148	0.81	3.15

*Rajgaria. R., McAllister. S. R., and Floudas C.A., Proteins, 70, 950-970 (2008)

*Tobi. D., and Elber. R., Proteins ,2000, 41, 40-46.

Results – Testing the Force Field

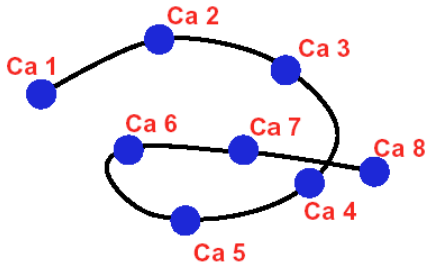
- Tested on five completely independent test sets from DecoysRUs* database
- Each set generated using a different algorithm
  - Success on one set does not guarantee success on others
- Test results also compared with Density Score Function (DSF)**

* *Samudrala, Levitt, Protein Science, 2000,9,1399-1401*

***Wang, Fain, Levitt, Samudrala, BMC Structurak Biology. 2004,4,8-26*

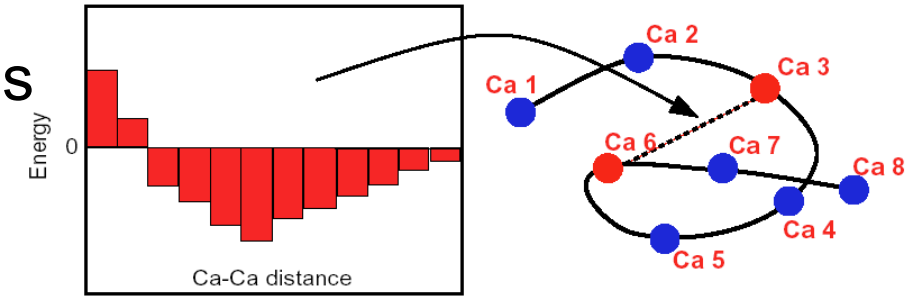
De Novo Protein Design Framework

Klepeis, Floudas, Lambris, Morikis 2003, 2004
Fung, Taylor, Floudas 2005, 2007, 2008



Sequence selection

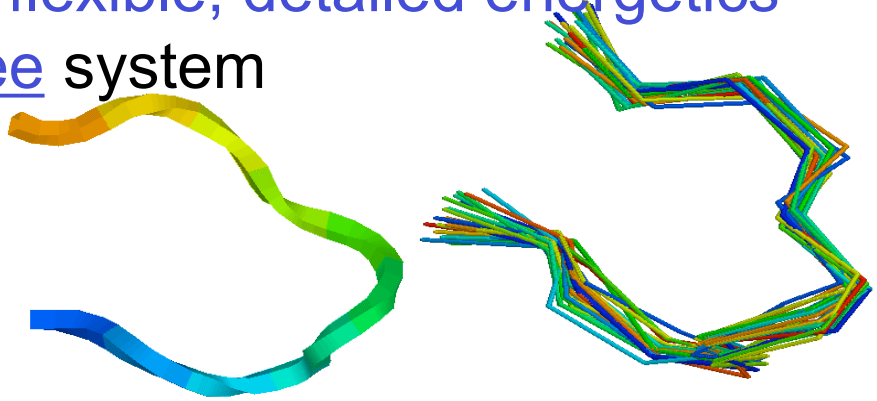
- Identify target template for desired fold; specify coordinates of backbone
- Identify possible residue mutations
- Introduce distance dependent pairwise potential based on Ca
- Generate ***rank-ordered list from mixed-integer linear (MILP)***



1	2	3	4	5	6	7	8
A	T	R	E	G	F	A	Q
A	S	K	E	P	Y	G	Q
V	S	K	E	G	F	A	Q

Fold Validation:Specificity

- Model selected sequences using flexible, detailed energetics
- Employ global optimization for free system
- Employ global optimization for system constrained to template
- Calculate relative probability for structures similar to desired fold



Sequence Selection : Key Ideas

- Consider template peptide of n positions
- At each position $i = 1,2,...,n$ there can be $j = 1,2,...,m_i$ mutations
- Define equivalent sets $k = 1,2,...,n$ and $l = 1,2,...,m_k$
- Require $k > i$ to represent all unique interactions
- Introduce 0-1 variables to indicate possible mutations at a given position

$$y_i^j = \begin{cases} 1 & \text{if residue type } j \text{ is in position } i \\ 0 & \text{otherwise} \end{cases}$$

$$y_k^l = \begin{cases} 1 & \text{if residue type } l \text{ is in position } k \\ 0 & \text{otherwise} \end{cases}$$

Mixed-integer Nonlinear Model

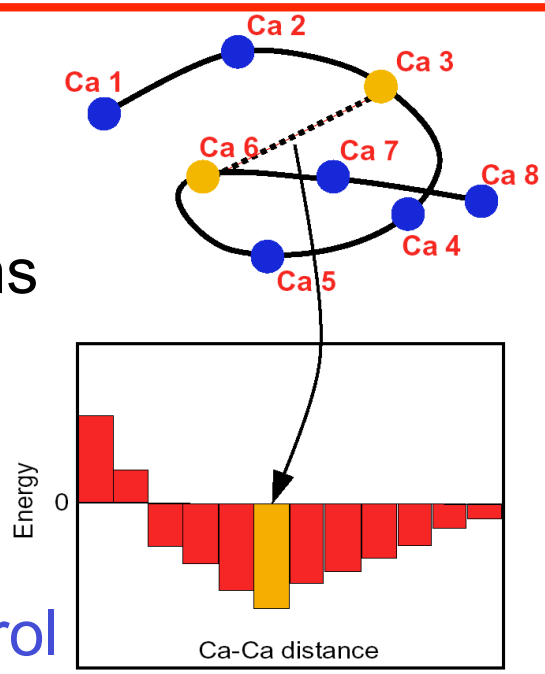
$$\min E(x) = \sum_{i=1}^{n}\sum_{j=1}^{m_i}\sum_{k>i}^{n}\sum_{l=1}^{m_k} E_{ik}^{jl}(x_i^j, x_k^l) y_i^j y_k^l$$

$$\text{subject to} \quad \sum_{j=1}^{m_i} y_i^j = 1 \qquad \forall \qquad i$$

$y_i^j = \begin{cases} 1 & \text{if residue type } j \text{ is in position } i \\ 0 & \text{otherwise} \end{cases}$

$y_k^l = \begin{cases} 1 & \text{if residue type } l \text{ is in position } k \\ 0 & \text{otherwise} \end{cases}$

- E_{ik}^{jl} is energy for protein with residue j at position I and residue l at position k
- taken from pairwise distance dependent energy function (x_{ij}, x_{kl}) using Ca positions
- parameters derived from MILP model to select native over low energy decoys



Important Remarks

- y_i^j and y_k^l are binary variables that control the residue type at a given position
- Binary variables appear bilinearly

→ Nonconvex

Mixed-integer Linear Reformulation

$$\min E(x) = \sum_{i=1}^{n}\sum_{j=1}^{m_i}\sum_{k>i}^{n}\sum_{l=1}^{m_k} E_{ik}^{jl}(x_i^j, x_k^l) w_{ik}^{jl}$$

$$\text{subject to} \quad \sum_{j=1}^{m_i} y_i^j = 1 \quad \forall \quad i$$

$$y_i^j + y_k^l - 1 \leq w_{ik}^{jl} \leq y_i^j \quad \forall \quad i,j,k,l$$

$$0 \leq w_{ik}^{jl} \leq y_k^l \quad \forall \quad i,k,j,l$$

$$\sum_{j=1}^{m_i} w_{ik}^{jl} = y_k^l \quad \forall \quad i,k,l$$

- Transform bilinear combinations to linear variables w_{ik}^{jl} Floudas 1995
- Reproduce properties of original formulation with constraints

  if $y_i^j = y_k^l = 0$ $w_{ik}^{jl} = 0$
  if $y_i^j = y_k^l = 1$ $w_{ik}^{jl} = 1$
  if y_i^j OR $y_k^l = 0$ $w_{ik}^{jl} = 0$

- Use Reformulation Linearization Technique (RLT) Sherali & coworkers based constraints to reduce integrality gap

Prove global optimality

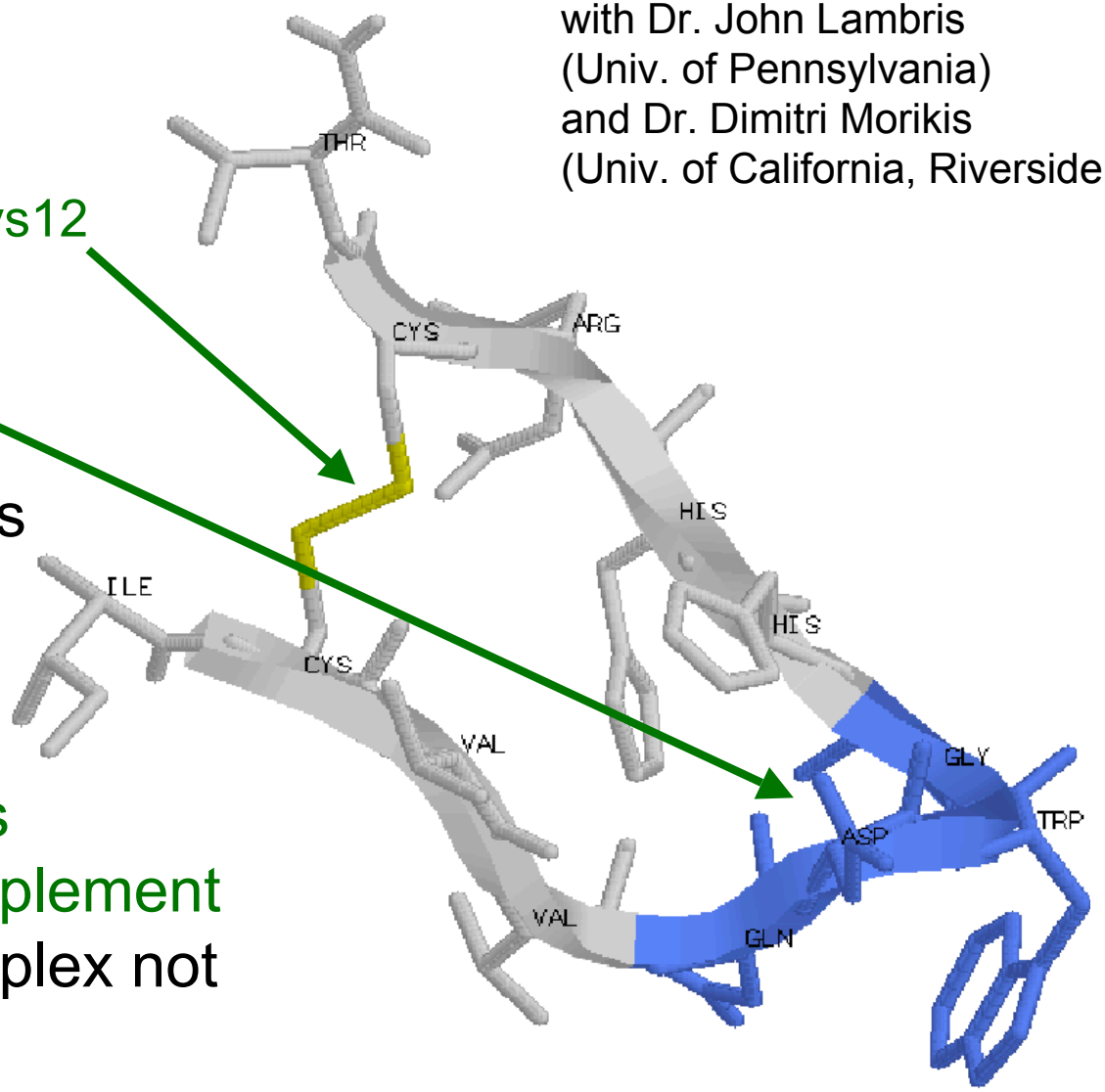
Compstatin

Potent inhibitor of third component of complement

with Dr. John Lambris
(Univ. of Pennsylvania)
and Dr. Dimitri Morikis
(Univ. of California, Riverside)

Structural features

- Cyclic, 13 residues
- Disulfide Bridge Cys2-Cys12
- Central beta-turn
  Gln5-Asp6-Trp7-Gly8
- Hydrophobic core
- Acetylated form displays higher inhibitory activity



Functional features

- Binds to and inactivates third component of complement
- Structure of bound complex not yet available

Sequence Selection : Compstatin

Design a more potent C3 inhibitor

Variable positions

- Conserve cystine residues (maintain cyclic nature of peptide)
- Conserve turn residues (do not overstabilize the turn)

Consensus results from top sequences

Position	Exp
1	A,V
4	Y,V,W
9	T,F,A
10	H
11	T,V,A,F,H
13	V,A,F

Key finding from computations

- His conserved at position 10
- Position 11 provides most variation : maintain Arg
- Selections at positions 4 and 9 allow for turn flexibility

New Enhancements of Quadratic Assignment like Models

- Three new algorithmic enhancement components to consider:

1. RLT with inequality constraints
2. Triangle inequalities
3. Preprocessing via DEE theorem

- Different combinations of the three components incorporated into the QA-like model to check for computational performance

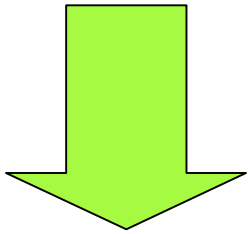
Fung, Rao, Floudas, Prokopyev, Pardalos, Rendl, J. Comb. Opt. (2005)

New Enhancements of Quadratic Assignment like Models_

- RLTs with inequality constraints

  Original RLT equations in Klepeis *et al*. (2004)'s formulation:

$$\sum_{j=1}^{m_i} w_{ik}^{jl} = y_k^l \qquad \forall \quad i,k,l$$



$$\sum_{j=1}^{m_i} w_{ik}^{jl} \le y_k^l \qquad \forall \quad i,k,l$$

i, k: alias sets for positions

j, l: alias sets for amino acids

"=" equals "≤" + "≥"

Implementing just "≤" should lead to a problem slightly easier and faster to solve

New Enhancements of Quadratic Assignment like Models_

- Triangle inequalities

$$(y_i^j - y_k^l)(y_i^j - y_m^p) \geq 0 \qquad \forall i < k < m, j, l, p$$

i,k,m: alias sets for positions

j,l,p: alias sets for amino acids

$$2 - (y_i^j - y_k^l)^2 - (y_i^j - y_m^p)^2 - (y_k^l - y_m^p)^2 \geq 0 \qquad \forall i < k < m, j, l, p$$

Valid inequalities that provide tighter lower bounds to the QA-like problem

Expand

Expand

$$y_i^j - w_{im}^{jp} - w_{ik}^{jl} + w_{km}^{lp} \geq 0 \qquad \forall i < k < m, j, l, p$$

$$w_{ik}^{jl} + w_{im}^{jp} + w_{km}^{lp} - y_i^j - y_k^l - y_m^p + 1 \geq 0 \qquad \forall i < k < m, j, l, p$$

★ Supposedly better to apply only a subset of the triangle inequalities by implementing: $S_{ikm}^{jlp} = E_{ik}^{jl} + E_{im}^{jp} + E_{km}^{lp} \leq$ arbitrary cutoff value

New Enhancements of Quadratic Assignment like Models_

- Preprocessing

  - Equation: $If \quad \exists \tilde{j} \neq j \quad s.t. \quad \sum_{k,k>i} \min_{l} [E_{ik}^{jl} - E_{ik}^{\tilde{j}l}] > 0$

    $then \quad y_i^j = 0$

  - Original idea came from the Dead-end Elimination criterion (Goldstein, 1994):

$$E(i_a) - E(i_b) + \sum_{k \neq i} \min_{c} [E(i_a, k_c) - E(i_b, k_c)] > 0$$

Rotamer i_a at position i can be pruned if its energy contribution is always lowered by substituting with an alternative rotamer i_b.

- We have only one “rotamer” for each amino acid

- Apply the DEE criterion for only one iteration

Stage one: New formulation for sequence selection

Fung, Floudas *et al.*, *J. Comb. Optim.*, 2005; Fung, Taylor, Floudas *et al.*, *OMS*, 2007

New sequence selection model for **single template structure**:

$$\min_{\substack{y_i^j, y_k^l \\ w_{ik}^{jl}}} \sum_{i=1}^{n} \sum_{j=1}^{m_i} \sum_{k=i+1}^{n} \sum_{l=1}^{m_k} E_{ik}^{jl}(x_i, x_k) w_{ik}^{jl}$$

$$\text{subject to} \quad \sum_{j=1}^{m_i} y_i^j = 1 \qquad \forall i$$

$$\sum_{j=1}^{m_i} w_{ik}^{jl} = y_k^l \qquad \forall i, k > i, l$$

$$\sum_{l=1}^{m_k} w_{ik}^{jl} = y_i^j \qquad \forall i, k > i, j$$

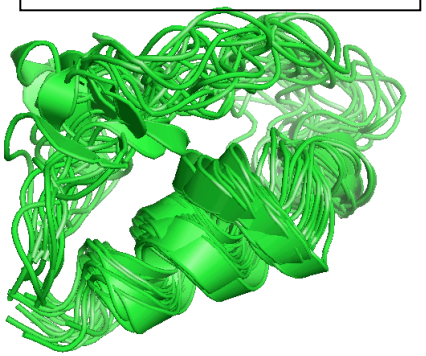
$$y_i^j, y_k^l, w_{ik}^{jl} = 0-1 \qquad \forall i, j, k > i, l$$

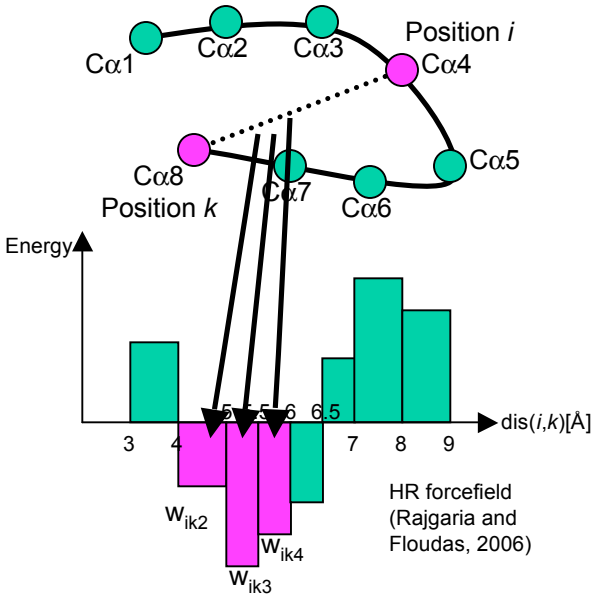
- obtained by declaring w_{ik}^{jl} as binary and new reduction properties
- we compared performance of the two models

Sequence selection models for flexible template with multiple structures: NEW MODELS

- We developed two different models:
  1. **Formulation using a weighted average of the structures**

Flexible design template





E_{ik}^{jl} $\sum_n w_{ikn} E_{ik}^{jl}$ in objective function

$$w_{ikn} = \frac{\text{no. of structures where dis}(i,k) \text{ falls into dist. bin } n}{\text{total no. of structures}}$$

$$\min_{y_i^j, y_k^l} \sum_{i=1}^{n} \sum_{j=1}^{m_i} \sum_{k=i+1}^{n} \sum_{l=1}^{m_k} \sum_{d=1}^{b_m} E_{ik}^{jl}(x_i, x_k) wt(x_i, x_k, d) w_{ik}^{jl}$$

subject to

$$\sum_{j=1}^{m_i} y_i^j = 1 \qquad \forall i$$

$$\sum_{j=1}^{m_i} w_{ik}^{jl} = y_k^l \qquad \forall i, k > i, l$$

$$\sum_{l=1}^{m_k} w_{ik}^{jl} = y_i^j \qquad \forall i, k > i, j$$

$$y_i^j, y_k^l, w_{ik}^{jl} = 0-1 \qquad \forall i, j, k, l$$

Sequence selection models for flexible template with multiple structures

2. Formulation using binary distance bin variables – Most General Model

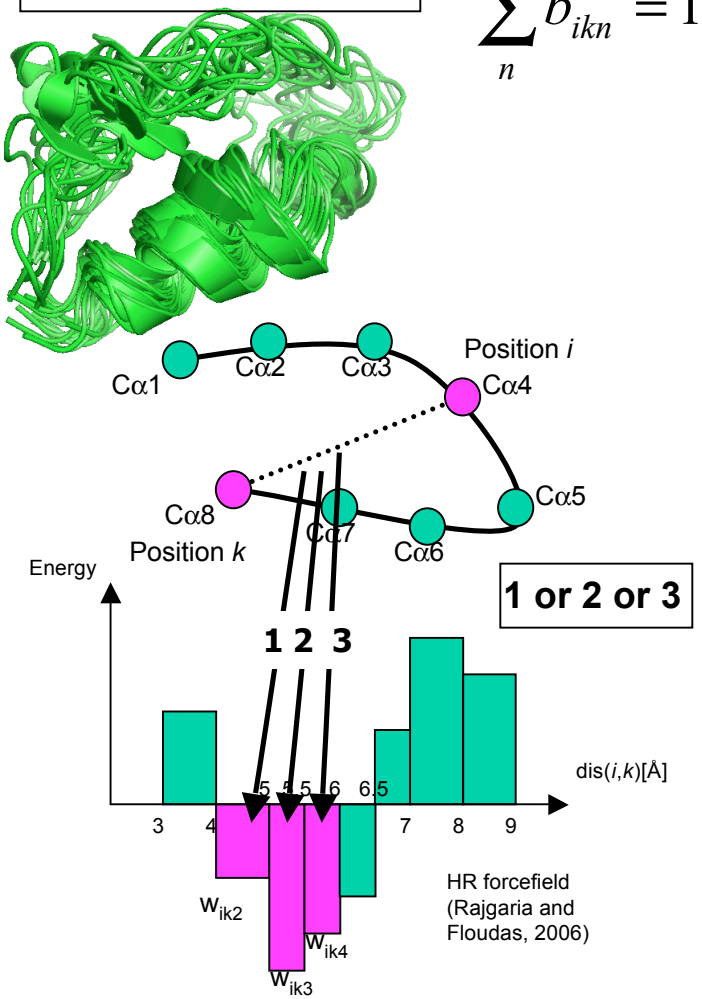
E_{ik}^{jl} changed to $\sum_n b_{ikn} E_{ik}^{jl}$ in objective function

$$\sum_n b_{ikn} = 1$$

b_{ikn} = 1 if dis(i,k) falls into dist. bin n

= 0 otherwise

Flexible design template



$$\min_{y_i^j, y_k^l} \sum_{i=1}^{n} \sum_{j=1}^{m_i} \sum_{k=i+1}^{n} \sum_{l=1}^{m_k} \sum_{d:disbin(x_i,x_k,d)=1} E_{ik}^{jl}(x_i,x_k) z_{ikd}^{jl}$$

subject to

$$\sum_{j=1}^{m_i} y_i^j = 1 \quad \forall i$$

$$\sum_{j=1}^{m_i} w_{ik}^{jl} = y_k^l \quad \forall i,k>i,l$$

$$\sum_{l=1}^{m_k} w_{ik}^{jl} = y_i^j \quad \forall i,k>i,j$$

$$\sum_{d:disbin(x_i,x_k,d)=1} b_{ikd} = 1 \quad \forall i,k>i$$

$$b_{ikd} + w_{ik}^{jl} - 1 \le z_{ikd}^{jl} \le b_{ikd} \quad \forall i,j,k>i,l,d$$

$$\sum_{d:disbin(x_i,x_k,d)=1} z_{ikd}^{jl} = w_{ik}^{jl} \quad \forall i,j,k>i,l$$

**

$$b_{ikd} + b_{kpd'} \le 1$$

if $(l_{mid}(d') < dis(i,p) - l_{mid}(d)$ or $l_{mid}(d') > dis(i,p) + l_{mid}(d))$

and $\sum_{d''=d+1}^{b_m} disbin(x_i,x_k,d'') \ge 1$ and $disbin(x_i,x_k,d) = 1$ and $disbin(x_k,x_p,d') = 1$

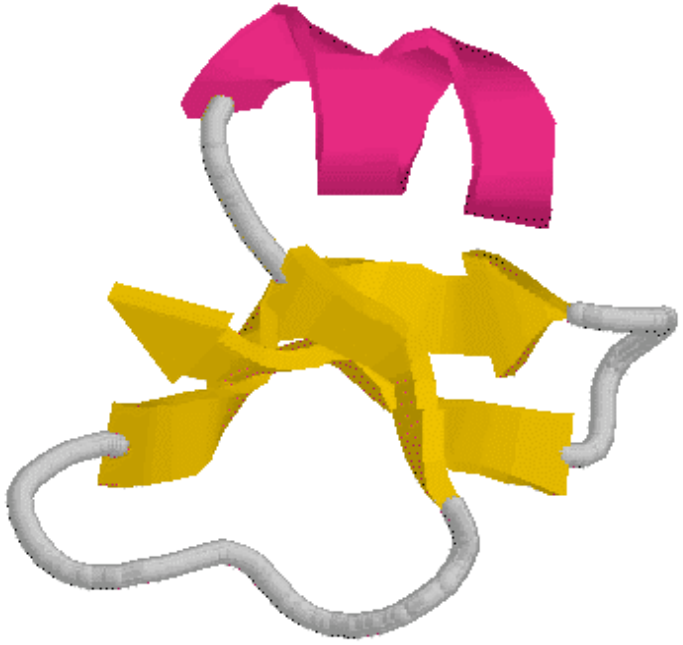
$$\forall i,k>i,p,d,d', i \ne k \ne p$$

$$y_i^j, y_k^l, w_{ik}^{jl}, b_{ikd}, b_{kpd'}, z_{ikd}^{jl} = 0\text{-}1 \; \forall i,j,k>i,l,p \ne k \ne i,d,d'$$

Antibacterial Peptides

Beta-Defensins

- Family of antimicrobial peptides
- Cationic peptides of 28-42 AAs
- Structure for only (2) humanBDs
  - Structure-function unknown
  - Low sequence identity
- hBD-2 10x more potent than hBD-1



	25					30					35					40				44
hBD-2	I	G	D	P	V	T	C	L	K	S	G	A	I	C	H	P	V	F	C	P
hBD-1	.	.	D	H	Y	N	C	V	S	S	G	G	Q	C	L	Y	S	A	C	P
mBD-7	.	N	S	K	R	A	C	Y	R	E	G	G	E	C	L	Q	.	R	C	I
mBD-8	.	N	E	P	V	S	C	I	R	N	G	G	I	C	Q	Y	.	R	C	I
hBD-3	T	L	Q	K	Y	Y	C	R	V	R	G	G	R	C	A	V	L	S	C	L
hBD-4	F	E	L	D	R	I	C	G	Y	G	T	A	R	C	R	K	.	K	C	R
mBD-1	.	.	D	Q	Y	K	C	L	Q	H	G	G	F	C	L	R	S	S	C	P
mBD-2	.	A	E	L	D	H	C	H	T	N	G	G	Y	C	V	R	A	I	C	P
mBD-3	I	N	N	P	V	S	C	L	R	K	G	G	R	C	W	N	.	R	C	I
mBD-4	I	N	N	P	I	T	C	M	T	N	G	A	I	C	W	G	.	P	C	P
bBD-1	.	.	D	F	A	S	C	H	T	N	G	G	I	C	L	P	N	R	C	P
bBD-2	.	.	N	H	V	T	C	R	I	N	R	G	F	C	V	P	I	R	C	P
bBD-12	.	.	G	P	L	S	C	G	R	N	G	G	V	C	I	P	I	R	C	P

	45					50					55					60				64
hBD-2	R	R	Y	K	Q	I	G	T	C	G	L	P	G	T	K	C	C	K	K	P
hBD-1	I	F	T	K	I	Q	G	T	C	Y	R	G	K	A	K	C	C	K	.	.
mBD-7	G	L	F	H	K	I	G	T	C	.	N	F	R	F	K	C	C	K	F	Q
mBD-8	G	L	R	H	K	I	G	T	C	.	G	S	P	F	K	C	C	K	.	.
hBD-3	P	K	E	E	Q	I	G	K	C	S	T	R	G	R	K	C	C	R	R	K
hBD-4	S	Q	E	Y	R	I	G	R	C	P	N	T	Y	A	.	C	C	L	R	K
mBD-1	S	N	T	K	L	Q	G	T	C	K	P	D	K	P	N	C	C	K	S	.
mBD-2	P	S	A	R	R	P	G	S	C	F	P	E	K	N	P	C	C	K	Y	M
mBD-3	G	N	T	R	Q	I	G	S	C	G	V	P	F	L	K	C	C	K	R	K
mBD-4	T	A	F	R	Q	I	G	N	C	G	H	F	K	V	R	C	C	K	I	R
bBD-1	G	H	M	I	Q	I	G	I	C	F	R	P	R	V	K	C	C	R	S	W
bBD-2	G	R	T	R	Q	I	G	T	C	F	G	P	R	I	K	C	C	R	S	W

Objective:

Design improved antibacterial peptides

De Novo Design of hβD-2

Structural features of Human beta defensin - 2:

Structural Feature	Positions
β Strands	14 - 16 25 - 28 36 - 39
α Helix	5 - 10
S-S bonds	8 - 37 15 - 30 20 - 38
β-Turns	16 - 19 21 - 24 32 - 35
Hairpins	25 - 29
Bulges	27, 28, 37

Constraints to add to model:

At least 2 hydrophobics on each β strand:

$$\sum_{i,j}(y_i^{Ala}+y_i^{Cys}+y_i^{Ile}+y_i^{Leu}+y_i^{Met}+y_i^{Phe}+y_i^{Trp}+y_i^{Tyr}+y_i^{Val})\geq 2 \qquad \forall 14\leq i\leq 16$$

$$\sum_{i,j}(y_i^{Ala}+y_i^{Cys}+y_i^{Ile}+y_i^{Leu}+y_i^{Met}+y_i^{Phe}+y_i^{Trp}+y_i^{Tyr}+y_i^{Val})\geq 2 \qquad \forall 25\leq i\leq 28$$

Quadratic Assignment Like Formulation Comparison

- Problem 2: full combinatorial optimization at pos. 2, 4, 6, 7, 9, 10, 11, 13, 14, and 16 (10 positions in total)

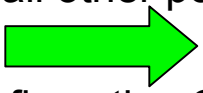
  all other pos. fixed at their native residues

  Sequence search space = 1.0×10^{13}

- Problem 3: full combinatorial optimization at pos. 2, 4, 6, 7, 9, 10, 11, 13, 14, 16, 18, 19, 22, 23, and 24 (15 positions in total)

  all other pos. fixed at their native residues

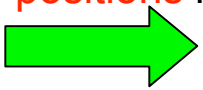
  Sequence search space = 3.3×10^{19}

- Problem 4: fix native Gly at pos. 1, 3, 12, 28, 31, and 34

  fix native Pro at pos. 5, 17, 21, 33, and 41

  fix native Cys at pos. 8, 15, 20, 30, 37, and 38

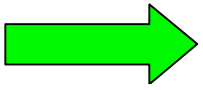
  full combinatorial optimization at all other positions (24 positions in total)

  Sequence search space = 1.7×10^{31}

- Problem 5: only fix native Cys at pos. 8, 15, 20, 30, 37, and 38

  full combinatorial optimization at all other positions (35 positions in total)

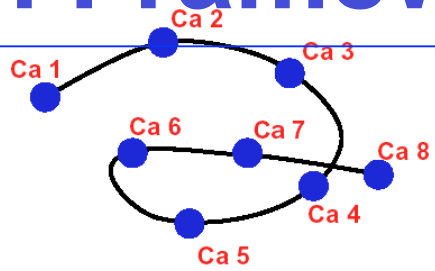
  Sequence search space = 3.4×10^{45}

Outline

- Motivational Examples
- De Novo Protein Design: Definition
- De Novo Design: Background
  - Advances and Limitations
- **Novel Two-Stage De Novo Protein Design Approach**
  - **Sequence Selection**
    - Force Fields: Distance Dependent
    - Quadratic Assignment-like Models
    - Compstatin, Human beta defensin-2
  - **Fold Specificity and Validation**
    - **First Principles based: Astro-Fold**
    - **NMR refinement like framework**
- Computational Studies
  - Design of Inhibitors for Complement 3
  - Design of C3a
  - Design of Inhibitors for HIV-1, gp41, gp120
- Conclusions & Acknowledgements

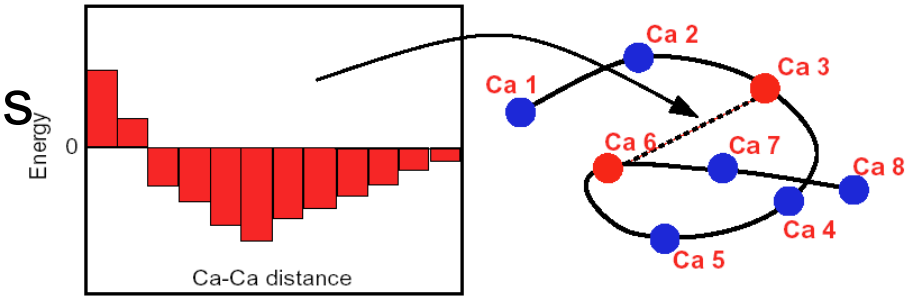
De Novo Protein Design Framework

Klepeis, Floudas, Lambris, Morikis 2003, 2004
Fung, Taylor, Floudas 2005, 2007, 2008



Sequence selection

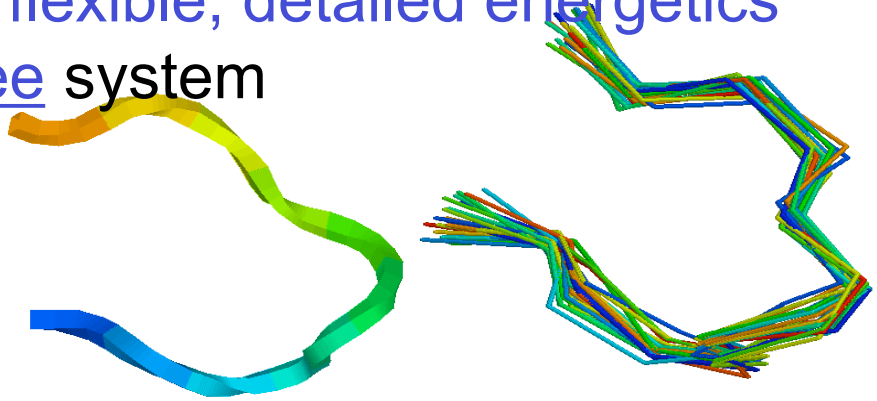
- Identify target template for desired fold; specify coordinates of backbone
- Identify possible residue mutations
- Introduce distance dependent pairwise potential based on Ca
- Generate rank-ordered energetic list from mixed-integer linear (MILP)



1	2	3	4	5	6	7	8
A	T	R	E	G	F	A	Q
A	S	K	E	P	Y	G	Q
V	S	K	E	G	F	A	Q

Fold Validation:Specificity

- Model selected sequences using flexible, detailed energetics
- Employ global optimization for free system
- Employ global optimization for system constrained to template
- Calculate relative probability for structures similar to desired fold



Fold Validation : Astro-Fold based

How to discriminate among the selected sequences

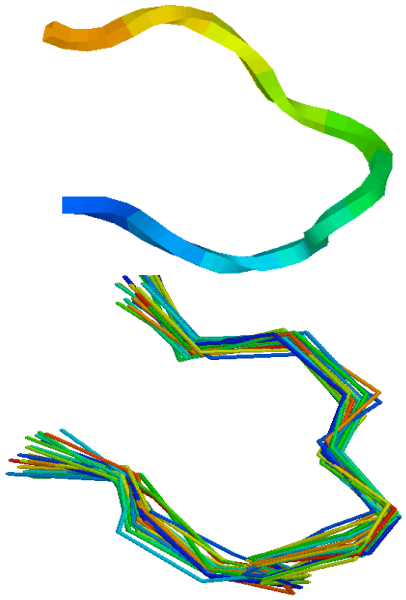
For each selected sequence solve (2) folding problems

- Free folding calculation

$$\min_{\boldsymbol{\theta}} \quad E(\boldsymbol{\theta})$$
$$\text{s.}t. \quad \text{Secondary structure constraints}$$

- Template constrained folding calculation

$$\min_{\boldsymbol{\theta}} \quad E(\boldsymbol{\theta})$$
$$\text{s.}t. \quad \text{Template + secondary constraints}$$



Quantify the specificity of the ensemble of structures similar to the template using probability calculation

$$p_{fold} = \frac{\sum_{i \in fold} \exp[-\beta(E_i)]}{\sum_{i \in (total)} \exp[-\beta(E_i)]}$$

Constrained Formulation

$$\begin{aligned} \min_{\boldsymbol{\theta}} \quad & E(\boldsymbol{\theta}) \\ \text{s.t.} \quad & E_{l,\text{distance}}(\boldsymbol{\theta}) \leq E_{l,\text{ref}} \quad l = 1,\ldots,N_{\text{con}} \\ & \theta_i^L \leq \theta_i \leq \theta_i^U \quad i = 1,\ldots,N_\theta \end{aligned}$$

Objective

- Nonconvex atomistic level force field

$$E = \sum_{ij \in NB} \epsilon_{ij} \left[\left(\frac{r_{ij}^o}{r_{ij}} \right)^{12} - \left(\frac{r_{ij}^o}{r_{ij}} \right)^{6} \right] + \sum_{ij \in HB} \epsilon_{ij} \left[\left(\frac{r_{ij}^o}{r_{ij}} \right)^{12} - \left(\frac{r_{ij}^o}{r_{ij}} \right)^{10} \right]$$
$$+ \sum_{ij \in ES} \frac{332\ q_i q_j}{D r_{ij}} + \sum_{k \in TOR} \frac{A_k}{2} (1 \pm \cos n_k \phi_k)$$

Constraints

- Enforce bounds on backbone variables
- Enforce upper / lower distances through square well constraints

$$E_{\text{distance}} = \sum_{j \in \text{upper}} \begin{cases} A_j (d_j - d_j^{\text{U}})^2 & \text{if } d_j > d_j^{\text{U}} \\ 0 & \text{otherwise} \end{cases}$$
$$+ \sum_{j \in \text{lower}} \begin{cases} A_j (d_j - d_j^{\text{L}})^2 & \text{if } d_j < d_j^{\text{L}} \\ 0 & \text{otherwise} \end{cases}$$

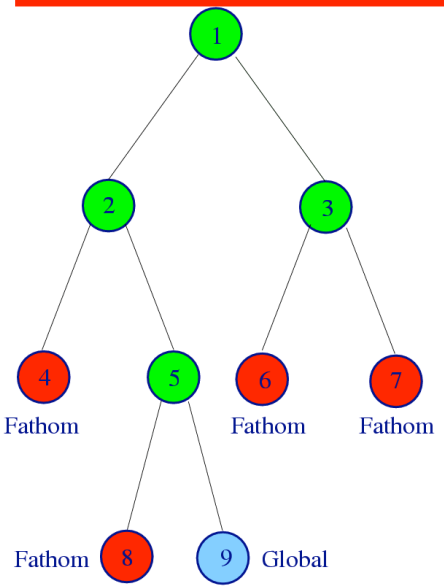
The αBB Framework

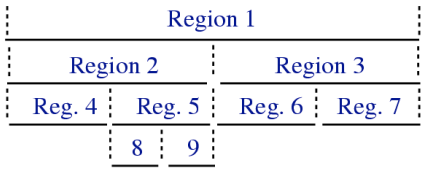
Floudas 2000
Floudas & co-workers
Adjiman et al. 1998,2001

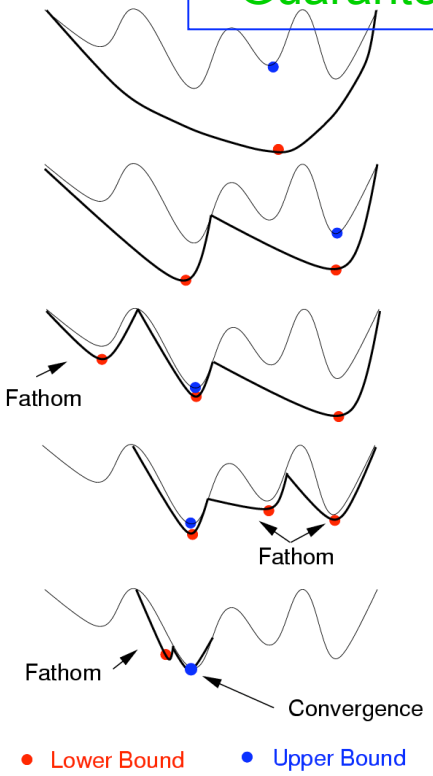
$$\begin{aligned} \min_{\mathbf{x}} \quad & f(\mathbf{x}) \\ \text{s.t.} \quad & \mathbf{h}(\mathbf{x}) = \mathbf{0} \\ & \mathbf{g}(\mathbf{x}) \leq \mathbf{0} \\ & \mathbf{x} \in \mathbf{X} \subseteq \mathcal{R}^n \end{aligned}$$

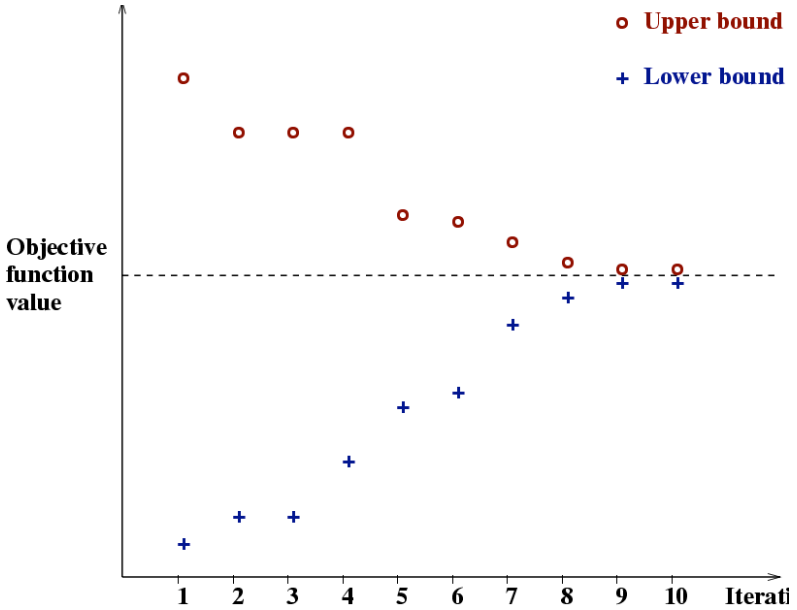
f , $\mathbf{h}$, $\mathbf{g}$ twice continuously differentiable

- Based on a branch-and-bound framework
- Upper bound on the global solution is obtained by solving the full nonconvex problem to local optimality
- Lower bound is determined by solving a valid convex underestimation of the original problem
- Convergence is obtained by successive subdivision of the region at each level in the brand & bound tree
- Guaranteed ε-convergence for C^2 NLPs









Torsion Angle Dynamics

Wuthrich & coworkers

Initialization

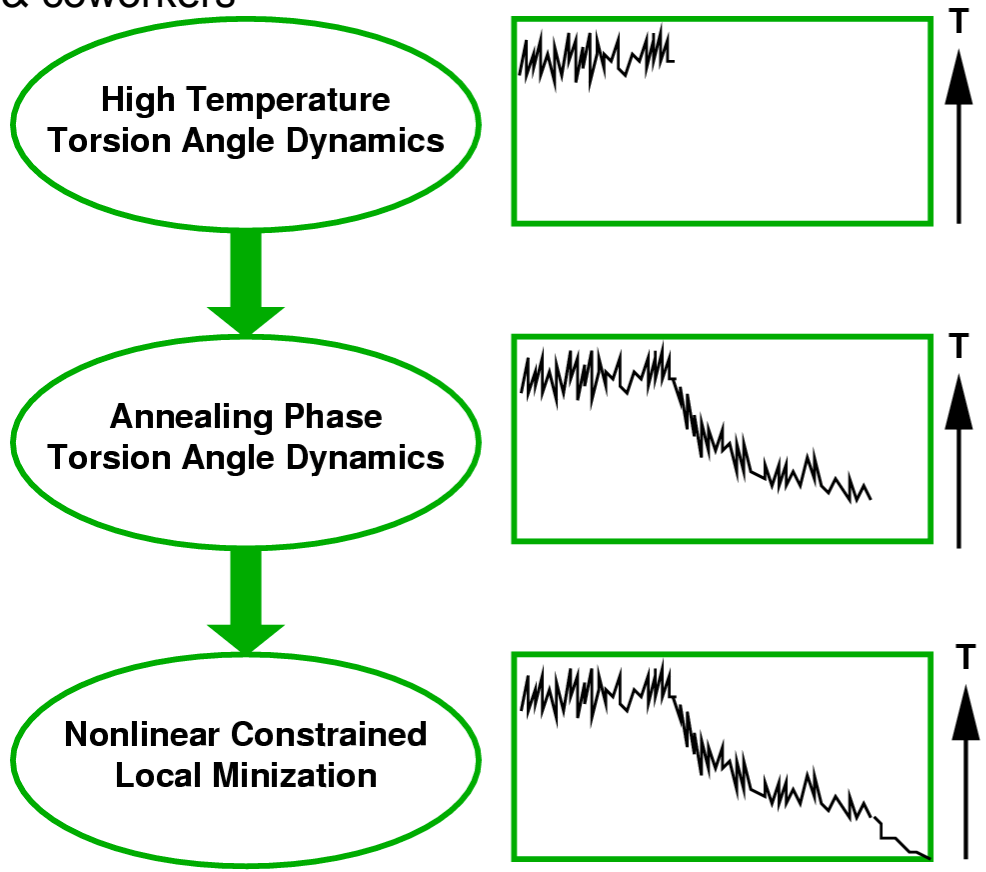
- Difficult to identify low energy feasible structures

Torsion Angle Dynamics

- Identify feasible low energy structures (satisfy constraints)
- Fast evaluation of simplified force field (steric based)
- Unconstrained formulation using penalty functions

Implementation

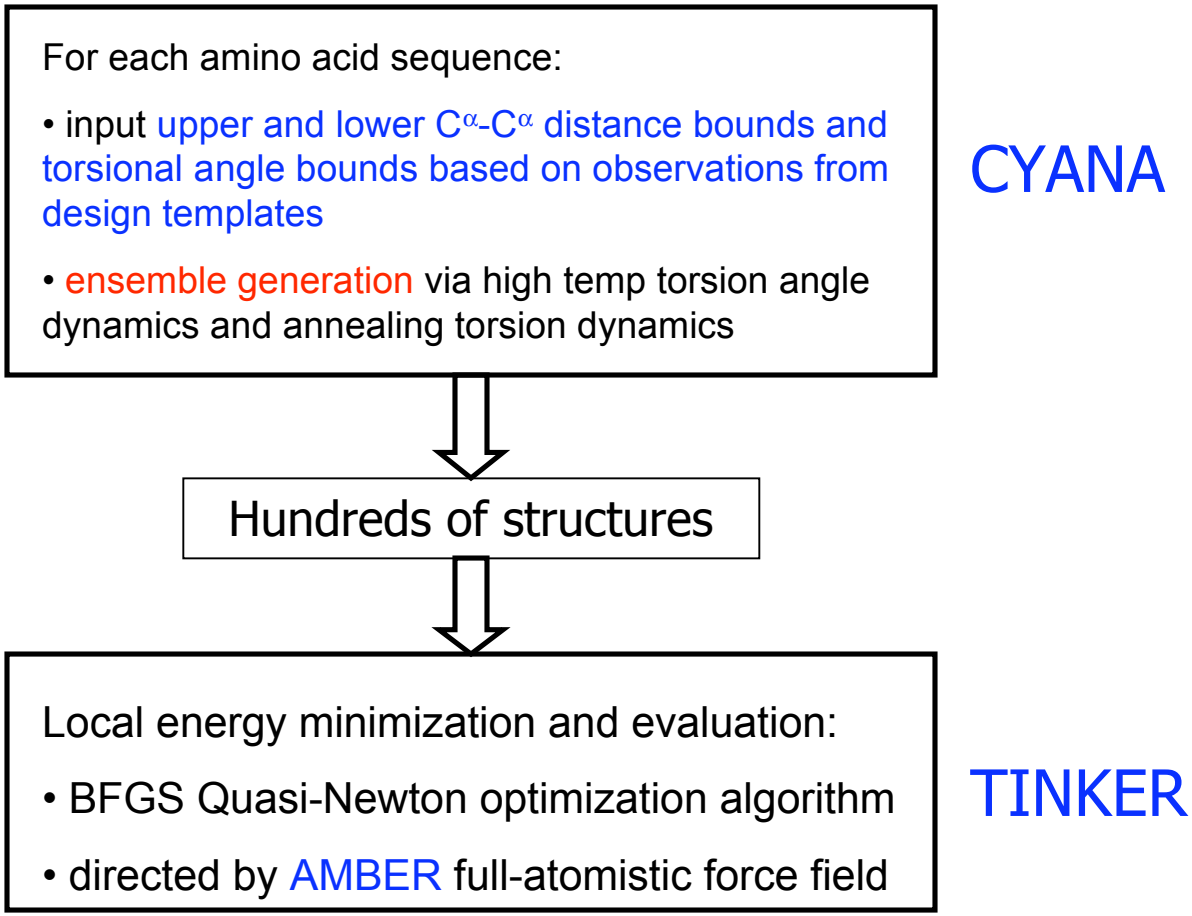
- Solve equations of motion as **preprocessing** for each constrained minimization



Good Initial Points for Local Optimization

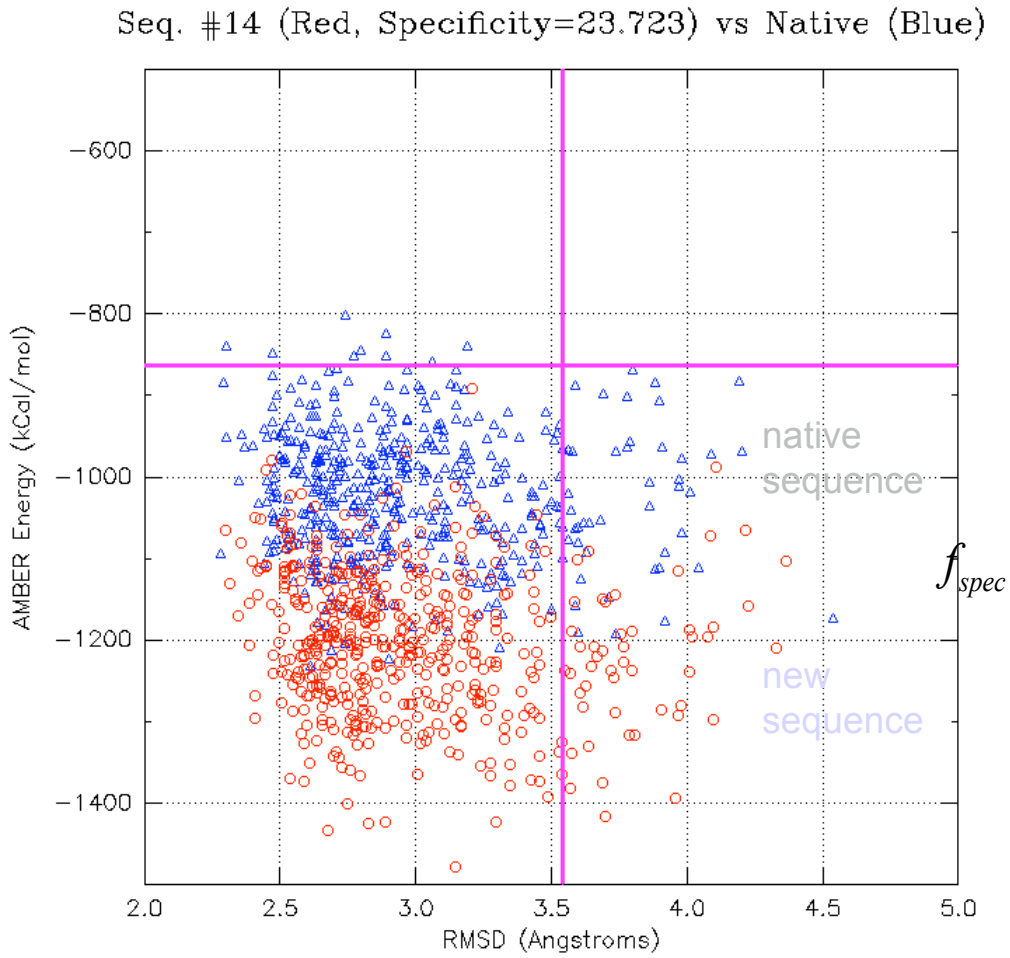
Fold Validation: NMR like framework

- Protein structure prediction method ASTRO-FOLD via first principles and deterministic global optimization: computationally expensive for large proteins (>200 residues)
- New fold specificity calculation method



Stage two: Fold Validation

- For each sequence from stage one, a specificity factor to the design template(s) is calculated



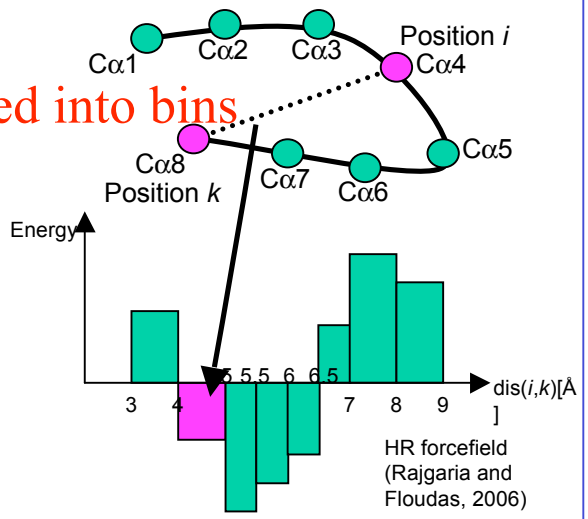
$$f_{spec} = \frac{\sum_{i \in \text{conformers of new sequence}} \exp(-\frac{E_i}{kT})}{\sum_{i \in \text{conformers of native sequence}} \exp(-\frac{E_i}{kT})}$$

FOLD SPECIFICITY CALC.

Framework allows for true backbone flexibility

- True backbone flexibility: bounded continuous distance and dihedral angles

- **Stage one**
  - distance dependence of energy is discretized into bins
  - models for flexible template with multiple structures & continuum



- **Stage two**
  - upper and lower bounds on distance and dihedral angles input by user
  - CYANA and TINKER-AMBER consider all possible combinations of continuous distance and angle values between bounds

Outline

- Motivational Examples
- De Novo Protein Design: Definition
- De Novo Design: Background
  - Advances and Limitations
- Novel Two-Stage De Novo Protein Design Approach
  - Sequence Selection
    - Force Fields: Distance Dependent
    - Quadratic Assignment-like Models
    - Compstatin, Human beta defensin-2
  - Fold Specificity and Validation
    - First Principles based: Astro-Fold
    - NMR refinement like framework
- **Computational Studies**
  - **Design of Inhibitors for Complement 3**
  - Design of C3a
  - Design of Inhibitors for HIV-1, gp41, gp120
- Conclusions & Acknowledgements

De Novo Design of Inhibitors for Complement 3: Compstatin variants

with Prof. J.D. Lambris (U. Penn) and

Prof. D. Morikis (UC, Riverside)

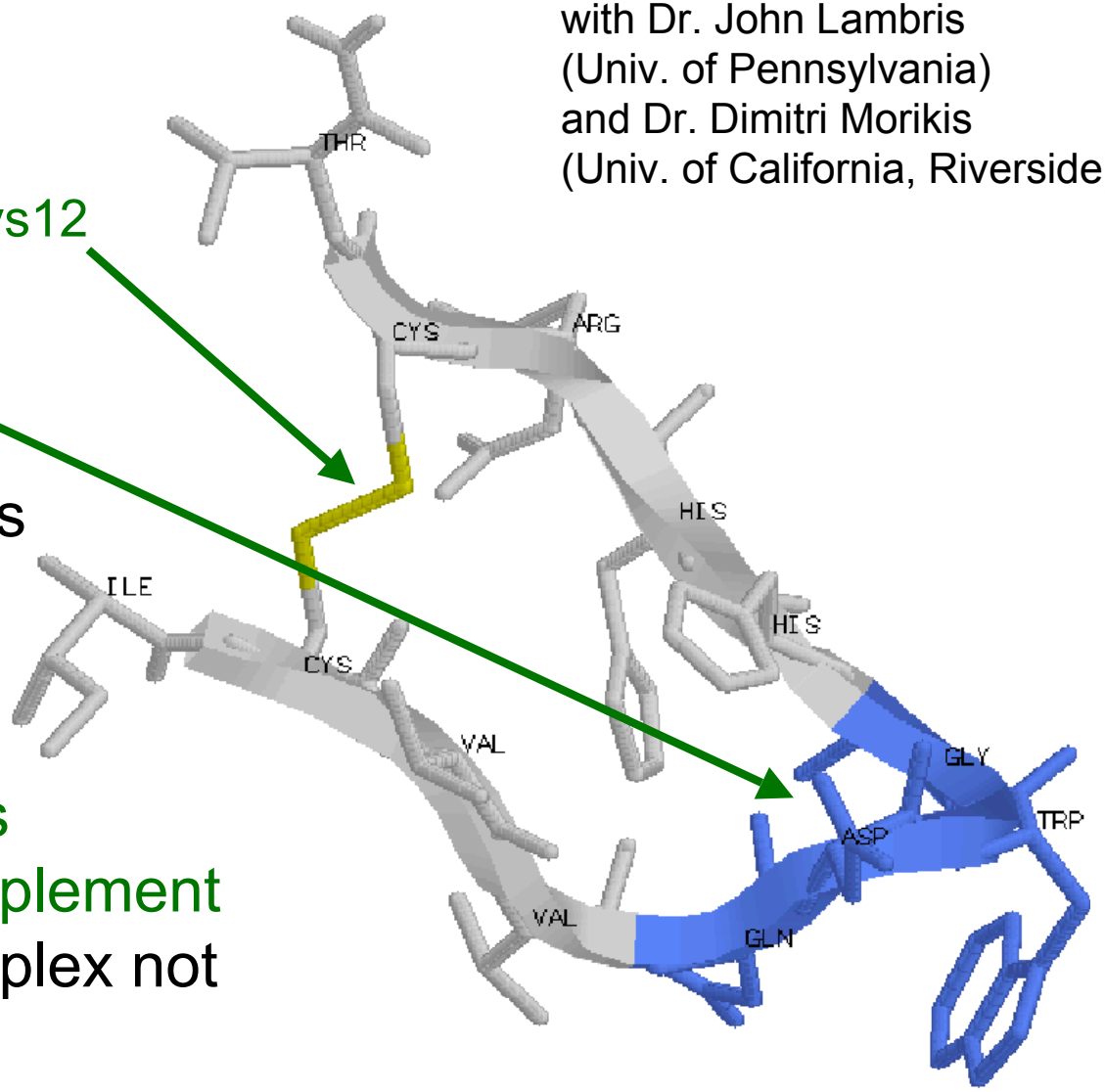
Compstatin

Potent inhibitor of third component of complement

with Dr. John Lambris
(Univ. of Pennsylvania)
and Dr. Dimitri Morikis
(Univ. of California, Riverside)

Structural features

- Cyclic, 13 residues
- Disulfide Bridge Cys2-Cys12
- Central beta-turn
  Gln5-Asp6-Trp7-Gly8
- Hydrophobic core
- Acetylated form displays higher inhibitory activity



Functional features

- Binds to and inactivates third component of complement
- Structure of bound complex not yet available

Sequence Selection : Compstatin

Design a more potent C3 inhibitor

Variable positions

- Conserve cystine residues (maintain cyclic nature of peptide)
- Conserve turn residues (do not overstabilize the turn)

Consensus results from top sequences

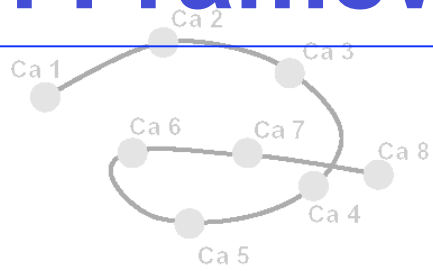
Position	Exp
1	A,V
4	Y,V,W
9	T,F,A
10	H
11	T,V,A,F,H
13	V,A,F

Key finding from computations

- His conserved at position 10
- Position 11 provides most variation : maintain Arg
- Selections at positions 4 and 9 allow for turn flexibility

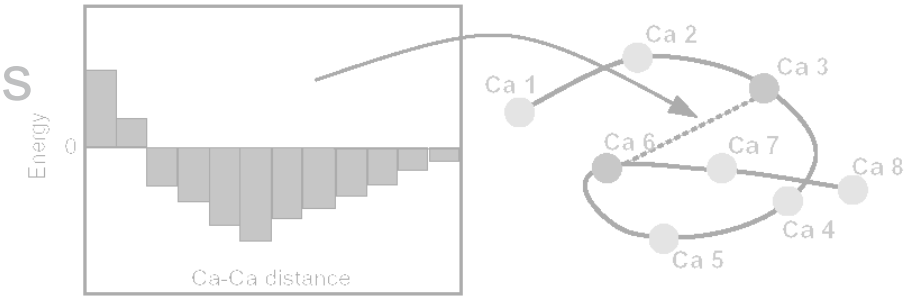
De Novo Protein Design Framework

Klepeis, Floudas, Lambris, Morikis 2003, 2004
Fung, Taylor, Floudas, 2005, 2007



Sequence selection

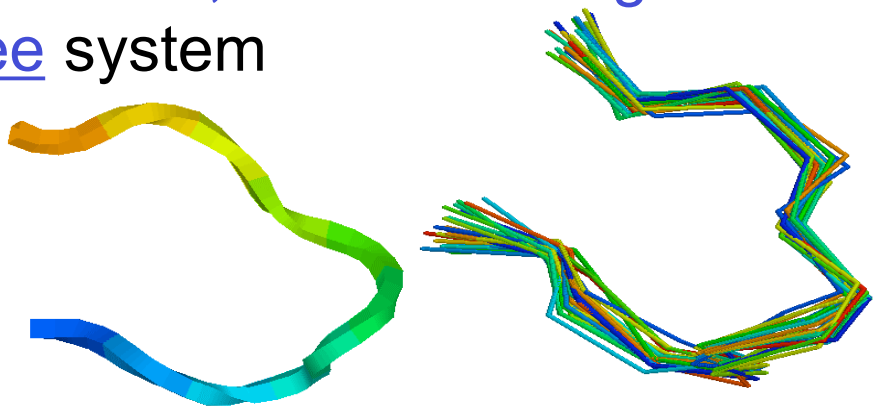
- Identify target template for desired fold; specify coordinates of backbone
- Identify possible residue mutations
- Introduce distance dependent pairwise potential based on Ca
- Generate rank-ordered energetic list from mixed-integer linear (MILP)



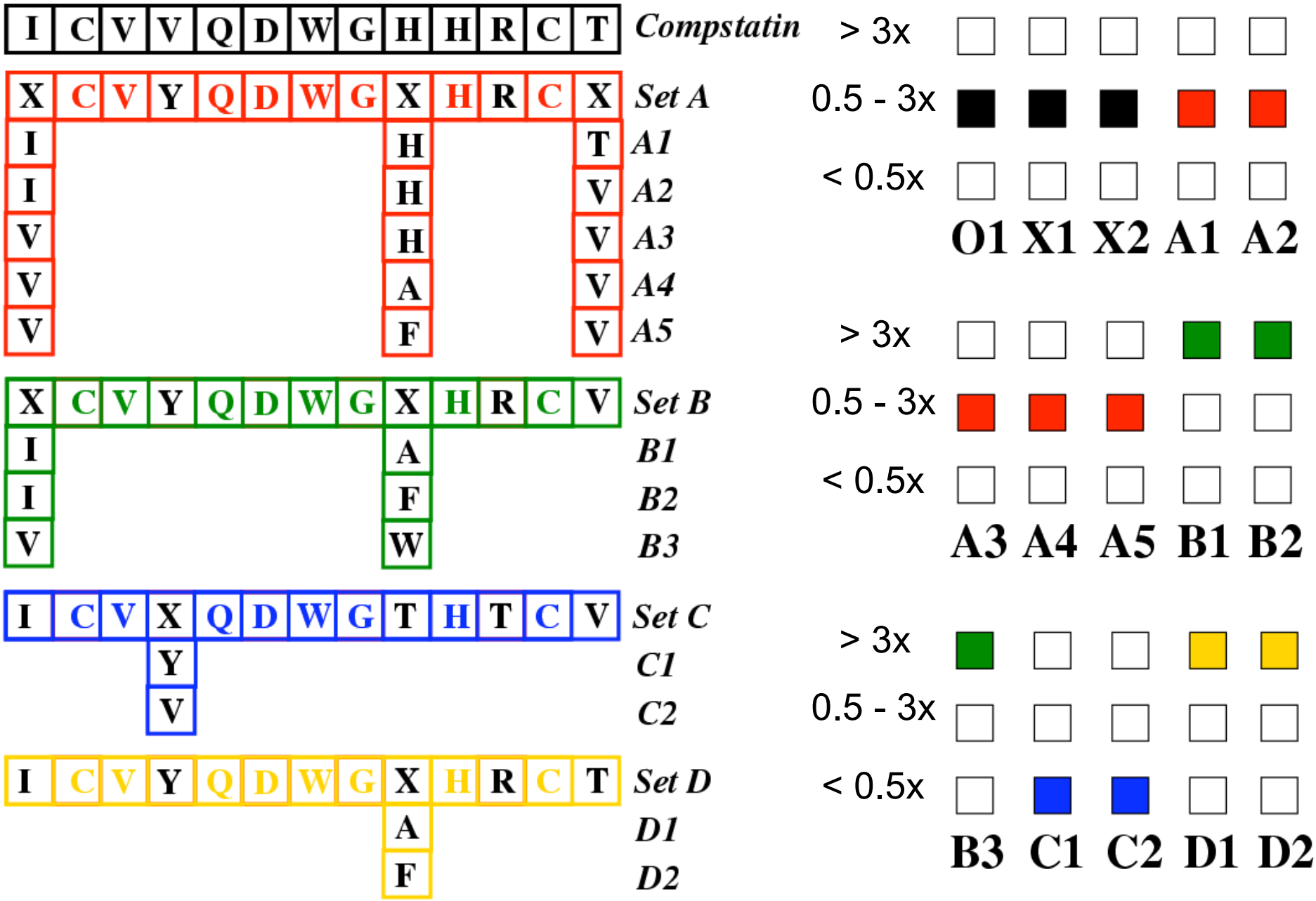
1	2	3	4	5	6	7	8
A	T	R	E	G	F	A	Q
A	S	K	E	P	Y	G	Q

Fold Validation

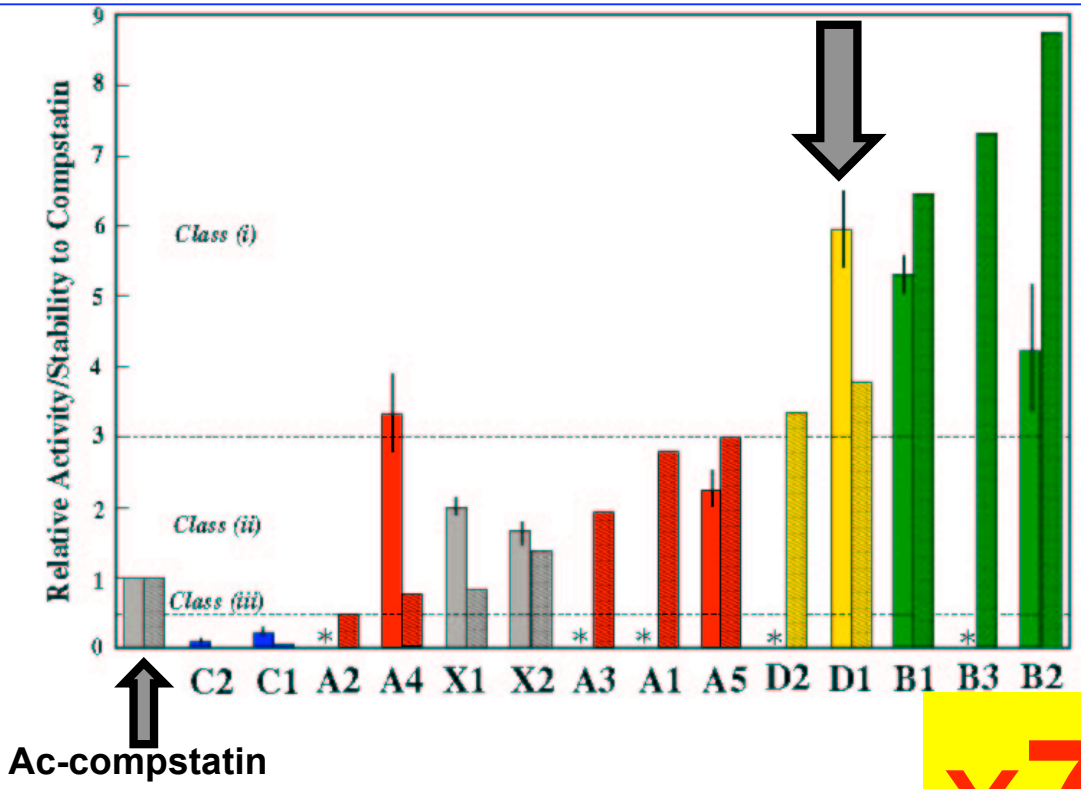
- Model selected sequences using flexible, detailed energetics
- Employ global optimization for free system
- Employ global optimization for system constrained to template
- Calculate relative probability for structures similar to desired fold



Compstatin Analogs



In Silico De Novo Design



Analog Ac-V4Y/H9A x7 x16

Analog Ac-W4Y/H9A x45

Klepeis, Floudas, Morikis, Tsokos, Argyropoulos, Spruce, Lambris (2003) J. American Chemical Society.

Klepeis, Floudas, Morikis, Lambris (2004) Ind. & Eng. Chem. Res.

Fung, Floudas (2005)

Outline

- Motivational Examples
- De Novo Protein Design: Definition
- De Novo Design: Background
  - Advances and Limitations
- Novel Two-Stage De Novo Protein Design Approach
  - Sequence Selection
    - Force Fields: Distance Dependent
    - Quadratic Assignment-like Models
    - Compstatin, Human beta defensin-2
  - Fold Specificity and Validation
    - First Principles based: Astro-Fold
    - NMR refinement like framework
- **Computational Studies**
  - Design of Inhibitors for Complement 3
  - **Design of C3a**
  - Design of Inhibitors for HIV-1, gp41, gp120
- Conclusions & Acknowledgements

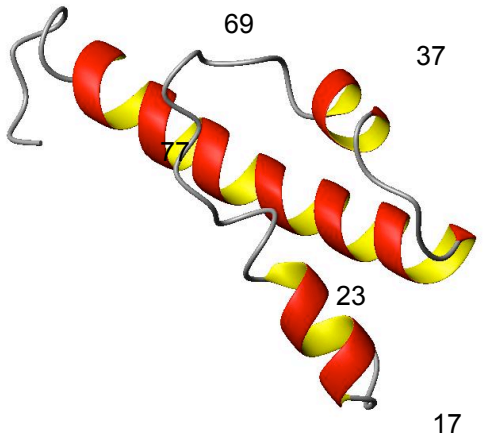
Redesign of Complement 3a

with Prof. D. Morikis (UC, Riverside)

Complement 3a

- Background:
  - fragment of the Complement 3 protein, **active mediator of inflammation**
  - 77-residue, 3 S-S bonds, 4 α-helices
  - sequence of C-terminal (pos 73-77) primary binding site: LGLAR
  - extensive sequence-activity studies by Ember *et al*. (1991)
  - super-potent peptide (12-15 times more active than natural C3a), WWGKKYRASKLGLAR (pos 63-77) identified by Ember *et al*. (1991)

- De novo design of C3a:
  - redesign pos 63-68, 70-72
  - ***Aim***: **identify peptides that are more active than natural C3a**

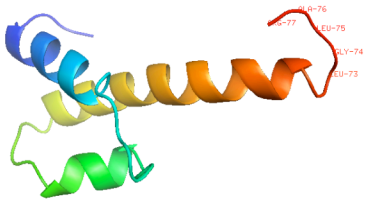


Residues 1-12 not shown

De novo design of C3a

- We used 3 sets of design templates:

1. Single structure from X-ray crystallography

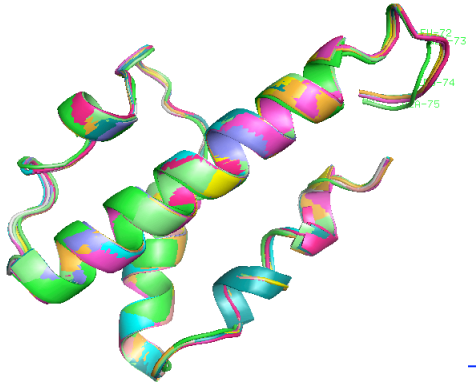


- Huber *et al.*, 1980
- Resolution: 3.2Å
- Residue 1 to 12 missing
- Side-chain information is also missing

2. Flexible templates from MD simulations with GB implicit solvation

- Initial structure: composite of C3a domain of C3's crystal structure (Janssen *et al.*, 2005) for Val^{1}-Ala^{70} and Huber *et al.*'s crystal structure for Ser^{71}-Arg^{77}
- **Starting from 10ns, one structure generated at each 1ns increment.**
- **11 flexible template structures in total**

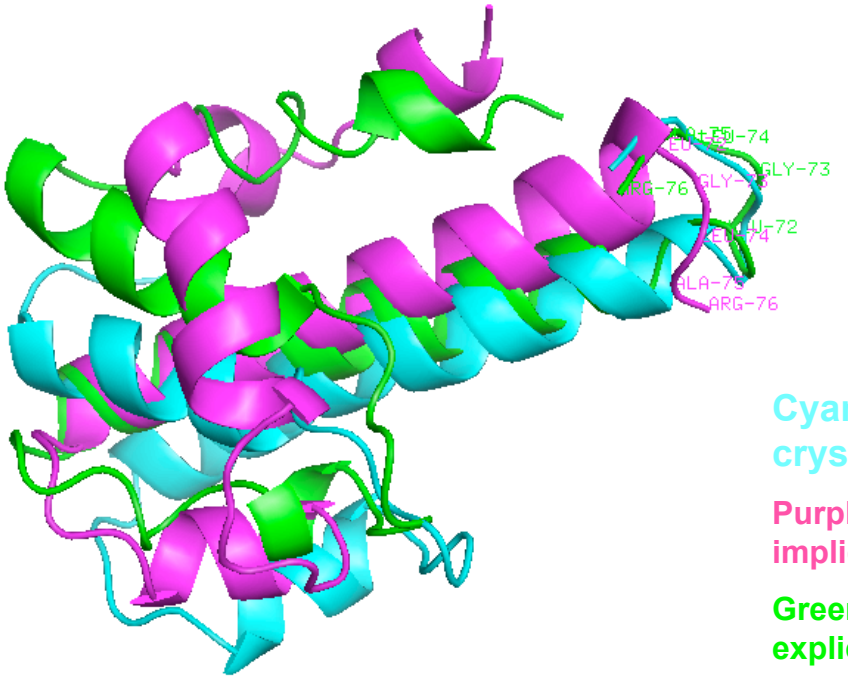
3. Flexible templates from MD simulations with explicit water molecules



- Structures generated using the same method as for the previous set of flexible templates except water molecules were treated explicitly in MD simulations
- **11 flexible template structures in total**

De novo design of C3a

- Structural deviation among the 3 sets of flexible design templates:



Cyan: single structure from X-ray crystallography

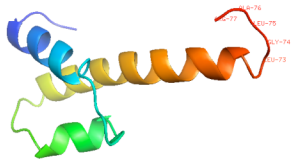
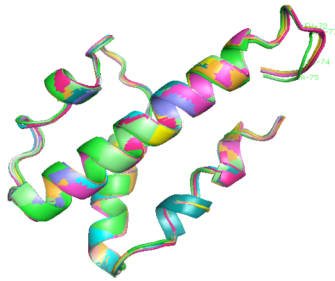
Purple: structure at 5 ns from MD simulations with GB implicit solvation

Green: structure at 5 ns from MD simulations with explicit water molecules

- The structural deviation means we should use all three sets of templates to get different predictions for active sequences

De novo design of C3a

- Forcefields and models applied for sequence selection:

Design templates	Forcefields	Sequence selection models
Single X-ray crystal structure 	• HR C^{α}-C^{α} forcefield only	• Basic model for single structure
MD simulations with GB implicit solvent 	• HR C^{α}-C^{α} forcefield • HR centroid-centroid forcefield	• Weighted average model for multiple structures • Binary distance bin model for multiple structures
MD simulations with explicit water 	• HR C^{α}-C^{α} forcefield • HR centroid-centroid forcefield	• Weighted average model for multiple structures • Binary distance bin model for multiple structures

- Generated 500 sequences for each

De novo design of C3a

- Mutation set for sequence selection

Position	63	64	65	66	67	68
SASA	54.6%	41.1%	51.6%	49.9%	31.0%	46.4%
Classification	surface	intermediate	surface	intermediate	intermediate	intermediate
Mutated?	yes	yes	yes	yes	yes	yes
Allowed residues	AILMFYWV	AILMFYWV RND QEGHKST	RNDQEGHKST	RND QEGHKST	RNDQEG HKST	AILMFYWVRND QEGHKST

69	70	71	72
51.3%	41.8%	36.4%	55.1%
surface	intermediate	intermediate	surface
no	yes	yes	yes
R	RNDQEGHKST A	RNDQEGHK ST	RNDQEG HKST

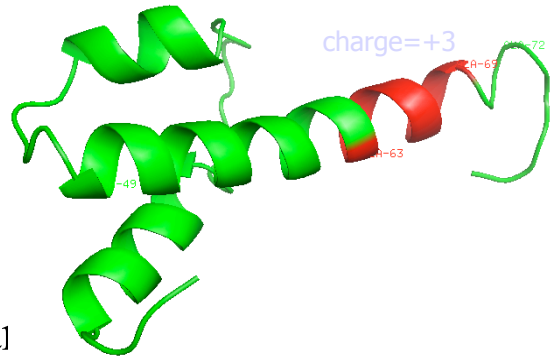
Problem complexity

$=2.59\times10^{9}$

- Biological constraints

- Maintain native charge on helix

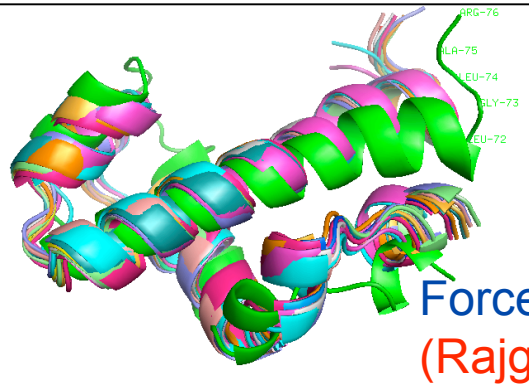
$$\sum_i y_i^{Arg} + \sum_i y_i^{Lys} - \sum_i y_i^{Asp} - \sum_i y_i^{Glu} = 3 \quad \forall\, 63 \le i \le 69$$



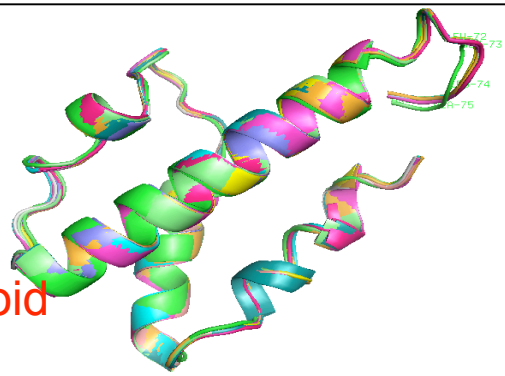
- Fold specificity stage: upper and lower bounds on a
based on observations about the flexible template(s).

Sequences from flexible templates generated with MD simulations

Design templates from MD simulations with GB implicit solvation



Design templates from MD simulations with explicit water molecules



Forcefield: HR centroid-centroid (Rajgaria et al., 2007)

Sequences synthesized																	flexible templates	sequence selection model
WT	L	R	R	Q	H	A	R	A	S	H	L	G	L	A	R		n. a.	n. a.
WR-15	W	W	R	S	K	W	R	E	E	Q	L	G	L	A	R		ex. water	binary dist. bin
WR-15-1	W	W	Q	R	R	W	R	D	E	Q	L	G	L	A	R		gen. Born	wt. avg.
WR-15-2	W	W	R	R	Q	W	R	D	E	Q	L	G	L	A	R		gen. Born	wt. avg.
WR-15-3	W	W	Q	R	R	W	R	D	E	R	L	G	L	A	R		gen. Born	wt. avg.
WR-15-4	W	W	Q	R	R	W	R	D	N	Q	L	G	L	A	R		gen. Born	wt. avg.
WR-15-5	W	W	Q	R	R	W	R	E	E	R	L	G	L	A	R		gen. Born	binary dist. bin
WR-15-6	W	W	R	R	Q	W	R	D	E	R	L	G	L	A	R		gen. Born	binary dist. bin
WR-15-7	W	W	R	R	Q	W	R	E	N	Q	L	G	L	A	R		gen. Born	binary dist. bin
WR-15-8	W	W	R	R	S	W	R	E	E	R	L	G	L	A	R		gen. Born	binary dist. bin
WR-15-9	W	W	R	N	R	W	R	E	N	R	L	G	L	A	R		gen. Born	binary dist. bin
WR-15-10	W	W	G	K	K	Y	R	A	S	K	L	G	L	A	R		n. a.	n. a.
WR-15-11	W	W	R	R	Q	W	R	E	D	H	L	G	L	A	R		ex. water	wt. avg.
WR-15-12	W	W	N	R	K	W	R	E	D	H	L	G	L	A	R		ex. water	wt. avg.
WR-15-13	W	W	R	R	Q	W	R	E	E	Q	L	G	L	A	R		ex. water	binary dist. bin
WR-15-15	W	W	R	R	Q	W	R	E	D	K	L	G	L	A	R		ex. water	binary dist. bin
WR-15-16	W	W	R	R	Q	W	R	E	E	H	L	G	L	A	R		ex. water	binary dist. bin
WR-15-17	W	W	R	R	H	W	R	E	D	Q	L	G	L	A	R		ex. water	binary dist. bin
WR-15-18	W	W	R	R	Q	W	R	E	E	K	L	G	L	A	R		ex. water	binary dist. bin
WR-15-19	W	W	R	R	Q	W	R	E	Q	K	L	G	L	A	R		ex. water	binary dist. bin

Super-agonist from Ember *et al.*, 1991

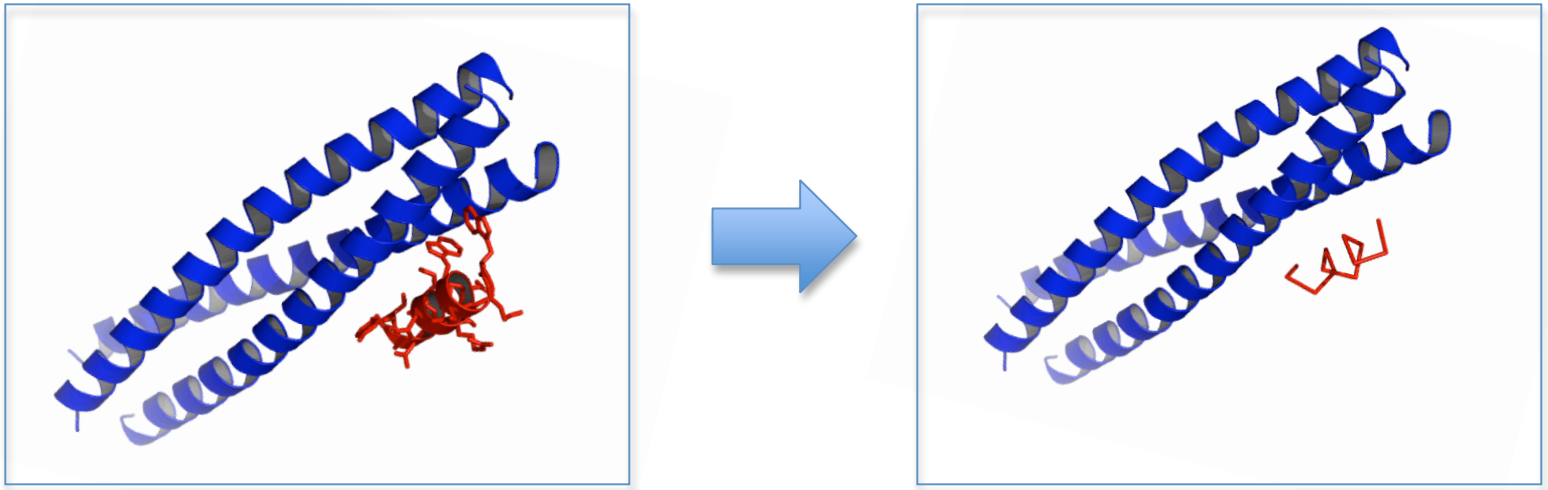
Outline

- Motivational Examples
- De Novo Protein Design: Definition
- De Novo Design: Background
  - Advances and Limitations
- Novel Two-Stage De Novo Protein Design Approach
  - Sequence Selection
    - Force Fields: Distance Dependent
    - Quadratic Assignment-like Models
    - Compstatin, Human beta defensin-2
  - Fold Specificity and Validation
    - First Principles based: Astro-Fold
    - NMR refinement like framework
- Computational Studies: Design of Inhibitors for Complement 3; Design of C3a
  - Approximate Binding Affinities for Protein-Ligand Interactions
- Computational Studies: Design of Inhibitors for C3c
  Design of Entry Inhibitors for HIV-1

Protein - Ligand Complexes: Approximate Binding Affinities

De Novo Design of Protein-Ligand Complexes

• Determination of **the amino acid sequences** that will fold into a certain **3-dimensional template structure** with higher stability or functionality than the native sequence



Previous Work

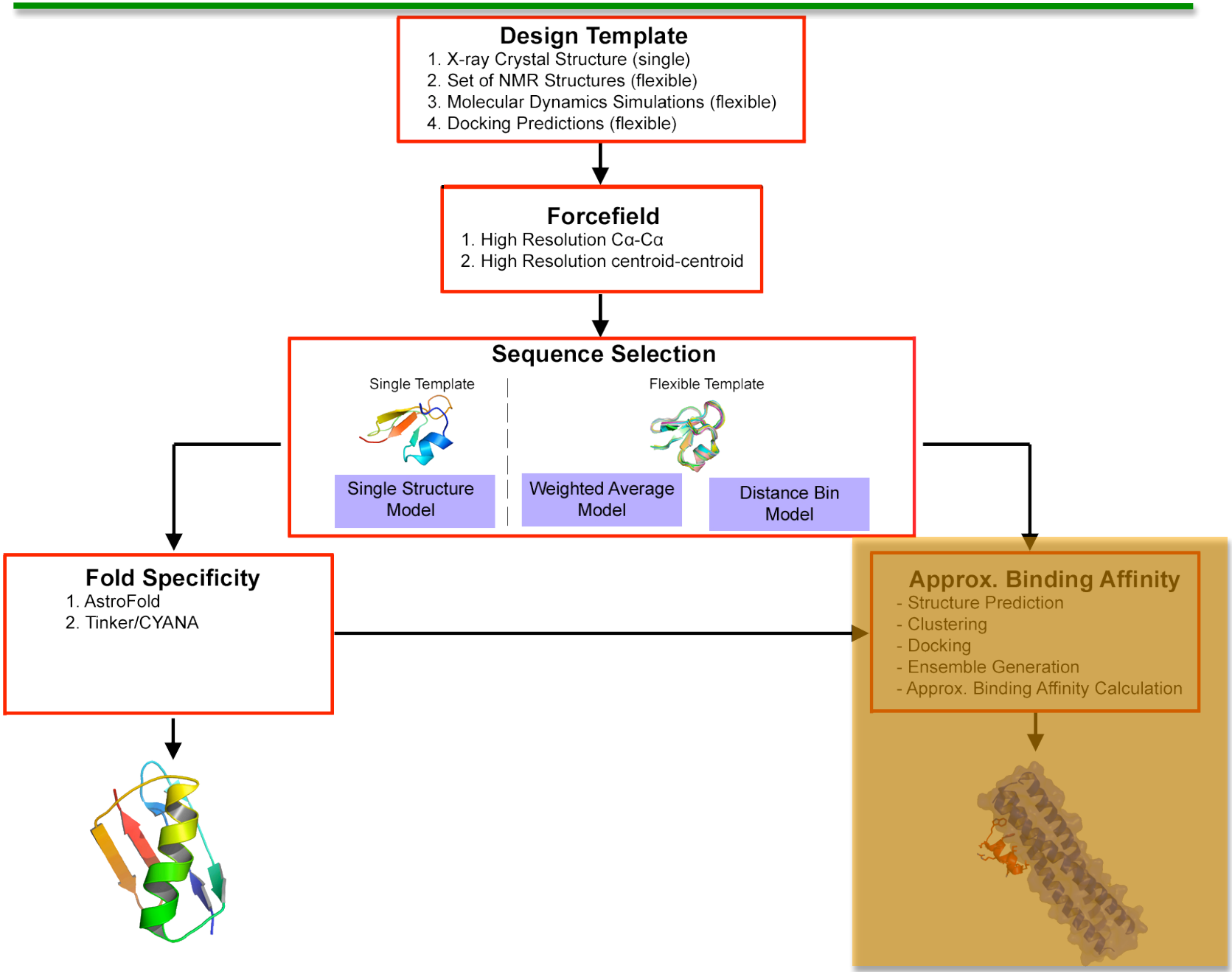
- **Rigid Templates**[1,2,3,4]
  - single fixed backbone
- **Flexible Templates**[5,6,7]
  - discrete rotamers on discrete templates
  - discrete rotamers on a continuum template
  - continuum template

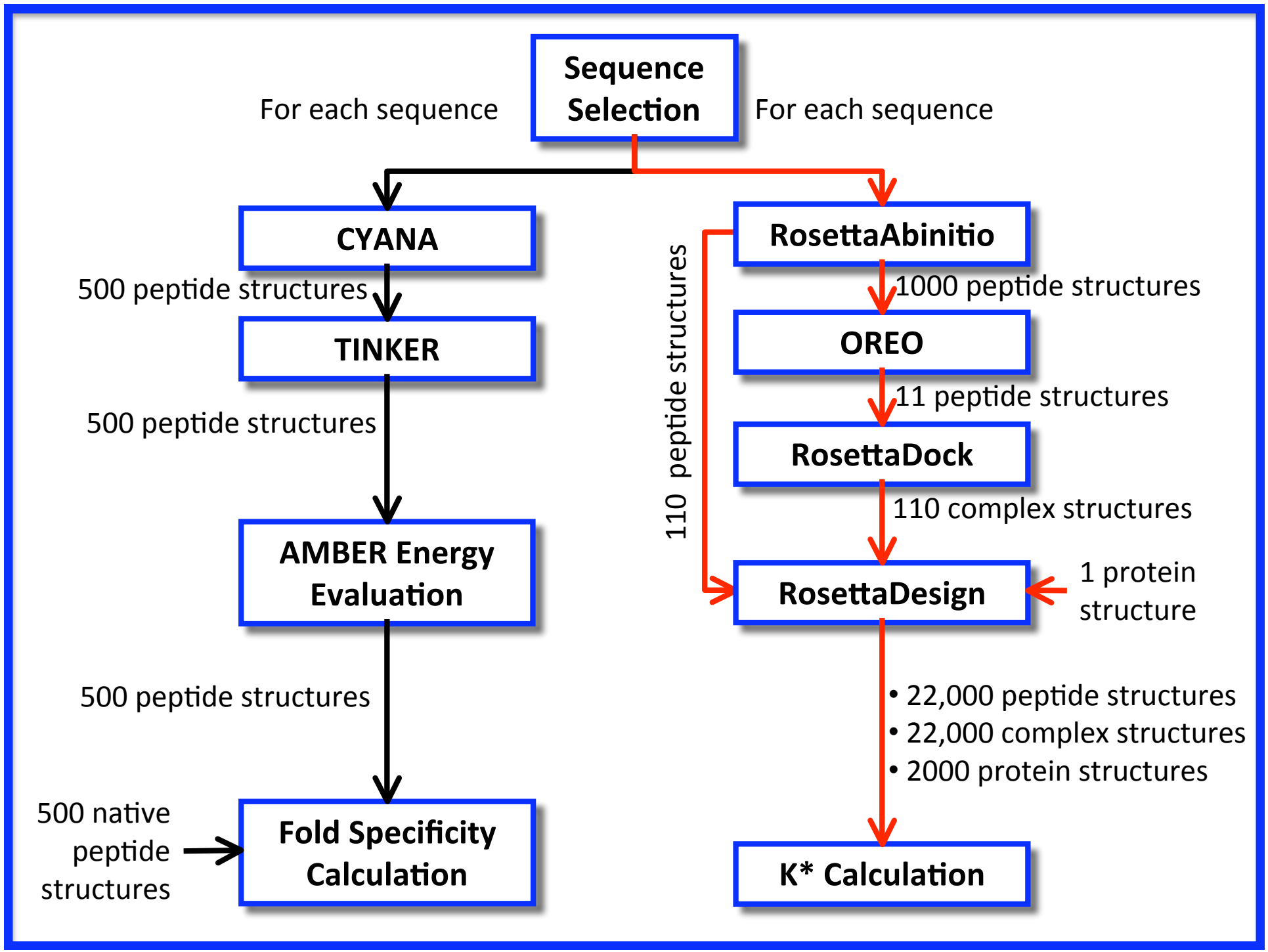
Design Model

- Flexible Template
- Mutation set
- Biological Constraints

[1]Desmet, J., *et al*. 1992 [2]Koehl, P. and M. Delarue 1994 [3]Dantas, G., *et al.* 2003 [4]Tuffery, P., *et al*. 1991
[5]Kono, H. and J. Saven 2001 [6]Harbury, P., *et al*. 1995 [7]Fung, H. K., *et al*. 2007

Protein Design Framework Overview




Sequence Selection
For each sequence
For each sequence
CYANA
500 peptide structures
TINKER
500 peptide structures
AMBER Energy Evaluation
500 peptide structures
500 native peptide structures
Fold Specificity Calculation
RosettaAbinitio
1000 peptide structures
OREO
11 peptide structures
RosettaDock
110 complex structures
110 peptide structures
RosettaDesign
1 protein structure
• 22,000 peptide structures
• 22,000 complex structures
• 2000 protein structures
K* Calculation

Ranking metric based upon approximate binding affinity calculations

$$K^* = \frac{q_{PL}}{q_P q_L} \qquad q_{PL} = \sum_{b \in B} e^{\frac{-E_b}{RT}}, \quad q_P = \sum_{f \in F} e^{\frac{-E_f}{RT}}, \quad q_L = \sum_{l \in L} e^{\frac{-E_l}{RT}}$$

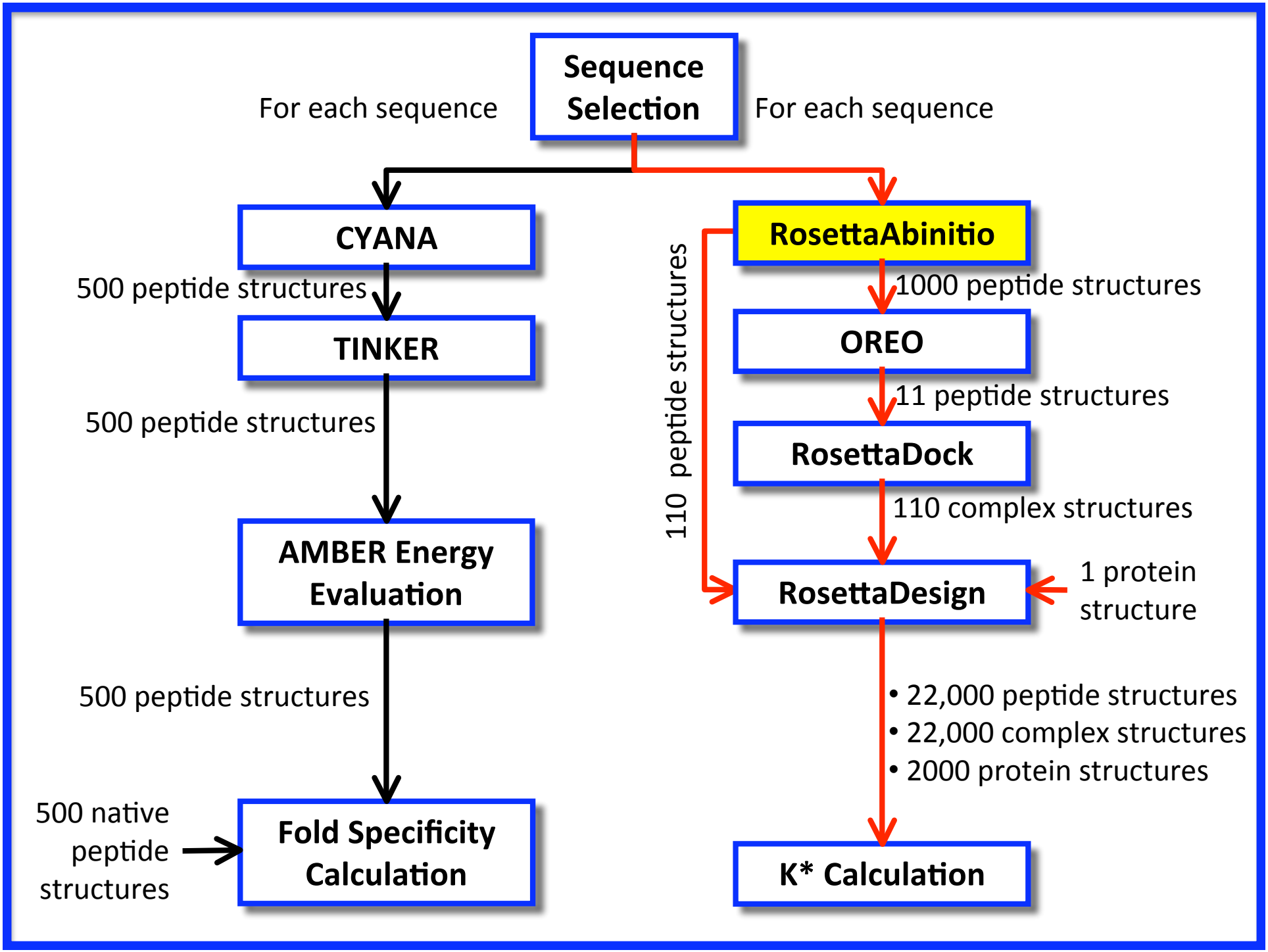
Derivation:

$$\mu_P + \mu_L + \mu_{PL} = 0 \quad (1)$$

$$\mu_i = -kT \ln \left(\frac{q_i(V,T)}{N_i} \right) \quad (2)$$

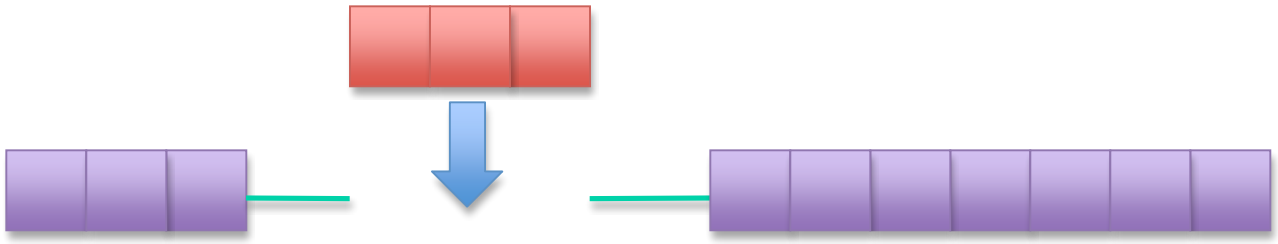
$$K^* = \frac{q_{PL}(V,T)}{q_P(V,T) q_L(V,T)} = \frac{N_{PL}}{N_P N_L} = \frac{[PL]}{[P][L]} = K_A \quad (3)$$

Lilien RH et al., J. of Computational Bio. (2005)

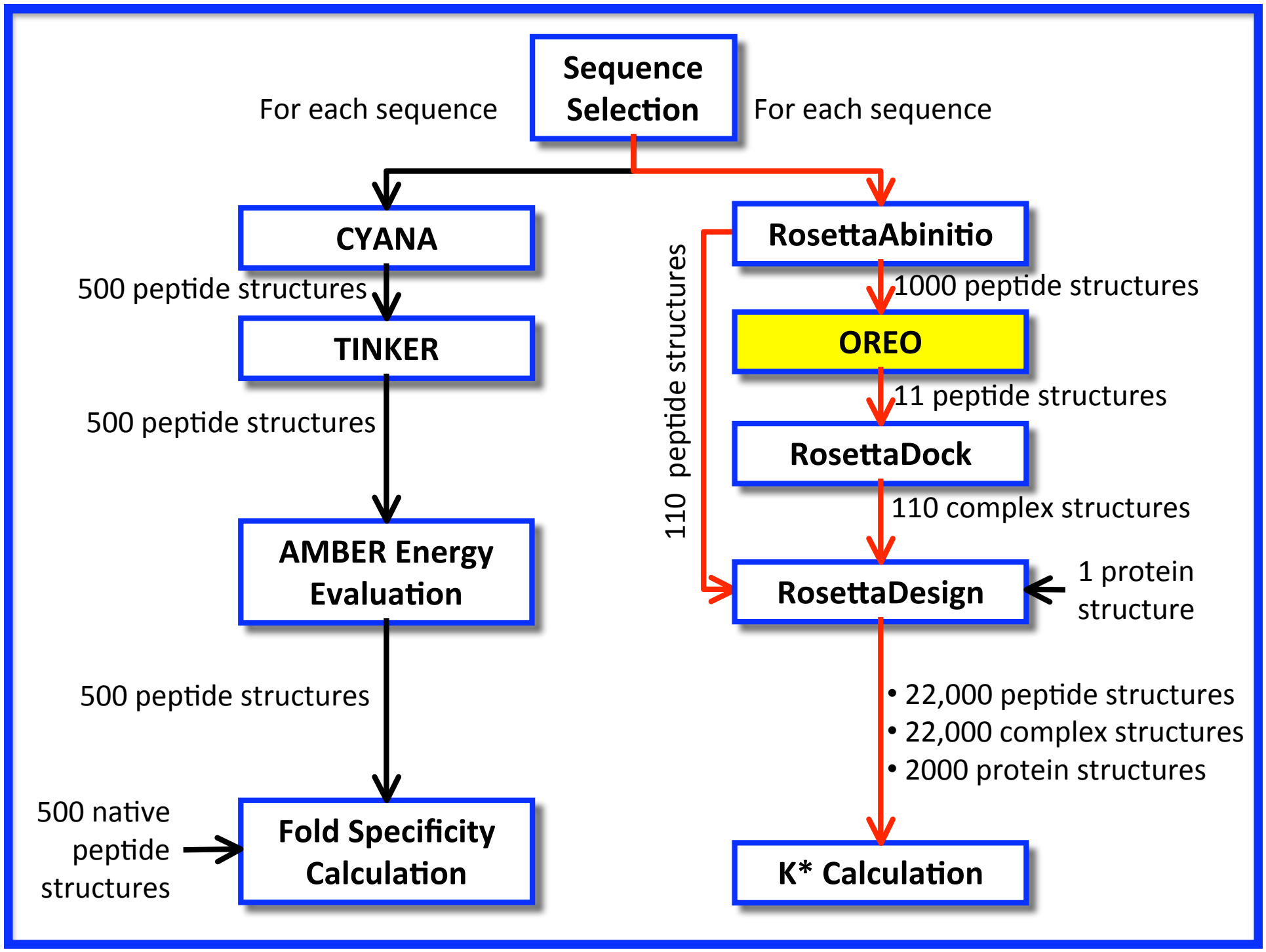

Sequence Selection
For each sequence
For each sequence
CYANA
500 peptide structures
TINKER
500 peptide structures
AMBER Energy Evaluation
500 peptide structures
500 native peptide structures
Fold Specificity Calculation
RosettaAbinitio
1000 peptide structures
OREO
11 peptide structures
RosettaDock
110 complex structures
110 peptide structures
RosettaDesign
1 protein structure
• 22,000 peptide structures
• 22,000 complex structures
• 2000 protein structures
K* Calculation

Peptide structure prediction using RosettaAbInitio

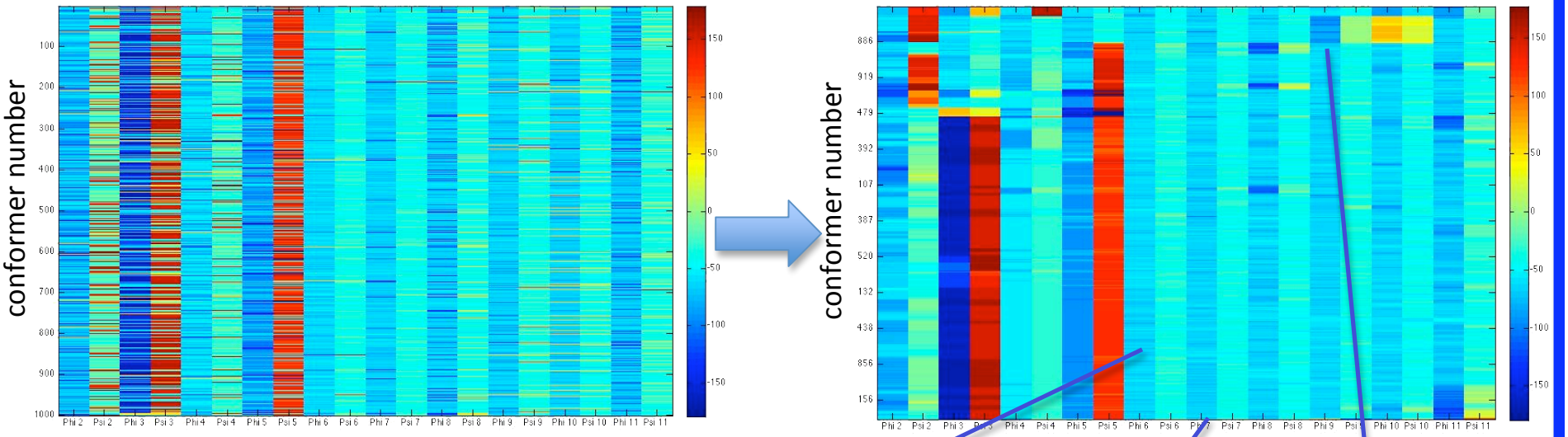
- Algorithm is based upon observation that **local structure** is **influenced**, but not uniquely determined, by **local sequence**
- Fragment Generation - **potential local structures**
- Fragment Insertion
  - Starting configuration is **extended chain**
  - Uses **Monte Carlo** search to randomly insert the fragments into the peptide chain
- Generate **1000 decoys**

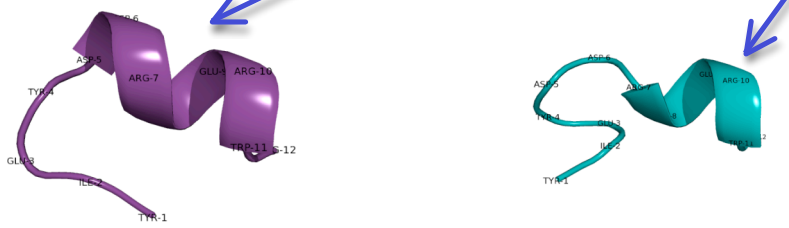


Rohl CA et al., Methods in Enzymology (2004)

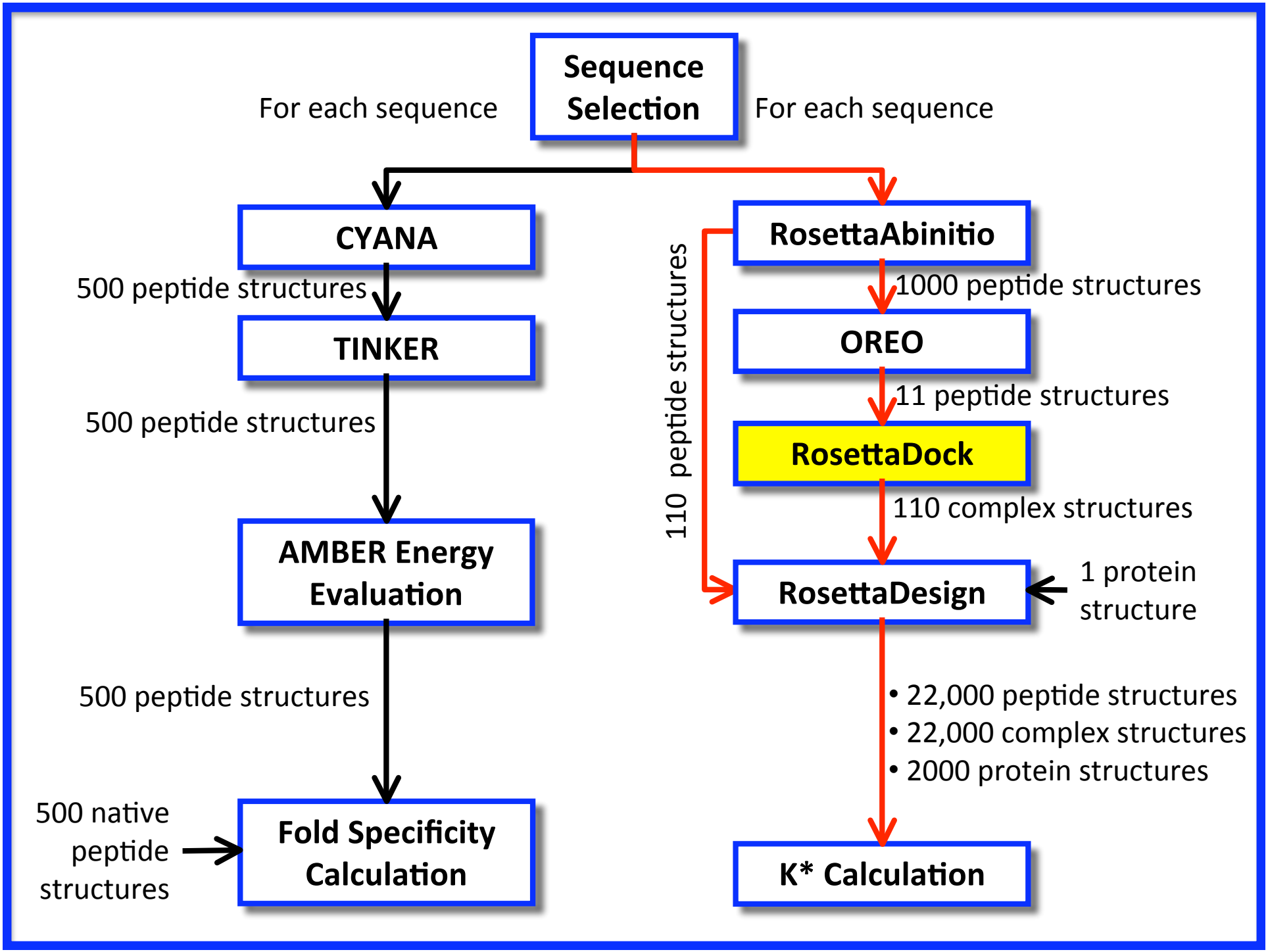

Sequence Selection
For each sequence
For each sequence
CYANA
500 peptide structures
TINKER
500 peptide structures
AMBER Energy Evaluation
500 peptide structures
500 native peptide structures
Fold Specificity Calculation
RosettaAbinitio
1000 peptide structures
OREO
11 peptide structures
RosettaDock
110 complex structures
110 peptide structures
RosettaDesign
1 protein structure
• 22,000 peptide structures
• 22,000 complex structures
• 2000 protein structures
K* Calculation

Clustering of peptide structures based on ϕ and ψ angles using OREO





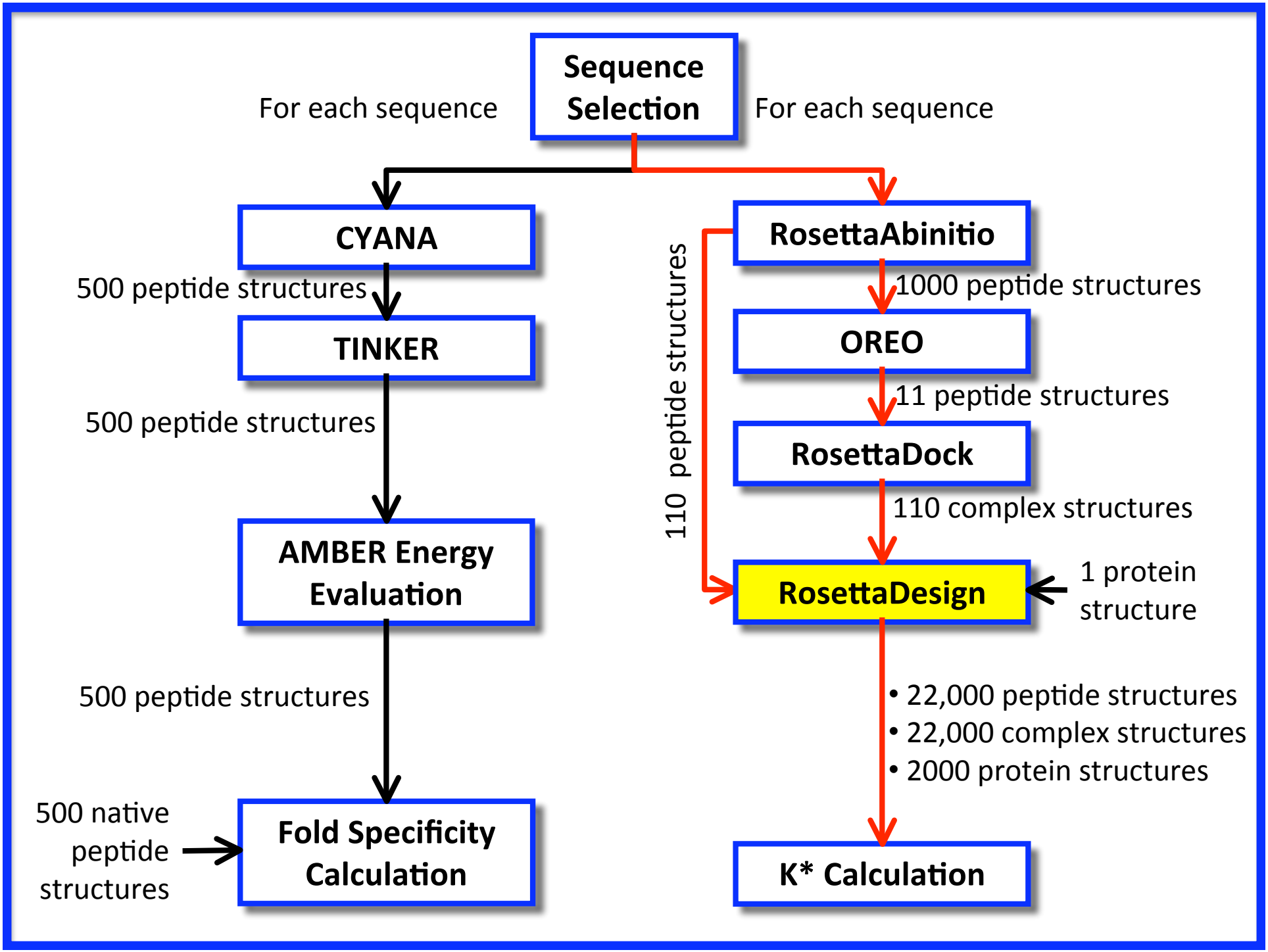
DiMaggio PA et al., BMC Bioinformatics (2008)
DiMaggio PA et al., J. of Global Optimization (2009)


Sequence Selection
For each sequence
For each sequence
CYANA
500 peptide structures
TINKER
500 peptide structures
AMBER Energy Evaluation
500 peptide structures
500 native peptide structures
Fold Specificity Calculation
RosettaAbinitio
1000 peptide structures
OREO
11 peptide structures
RosettaDock
110 complex structures
110 peptide structures
RosettaDesign
1 protein structure
• 22,000 peptide structures
• 22,000 complex structures
• 2000 protein structures
K* Calculation

Docking simulations of lowest energy peptide and top 10 Cluster medoids using RosettaDock

- Most **computationally intensive** step
- Rosetta Dock also uses a **Monte Carlo search** to translate and rotate one docking partner against the other
- Generate **1000 decoys** for each peptide-protein complex (11 peptides × 1000 decoys = 11000 total structures)
- Extract **10 lowest** scoring decoys from each docking simulation for ensemble generation starting structures (110 starting structures total)

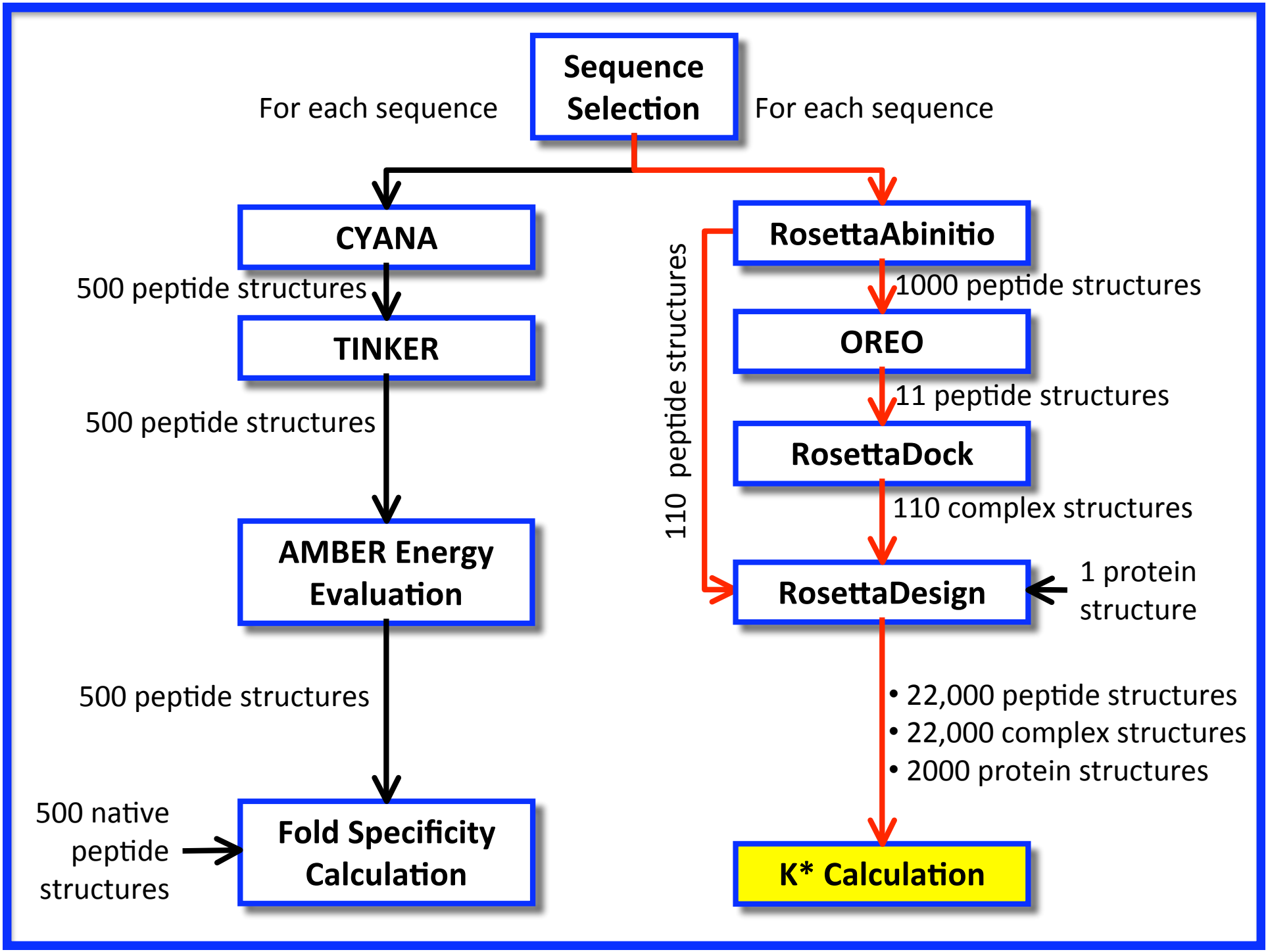
Gray JJ et al., J. of Mol. Bio. (2003)
Gray JJ et al., Proteins: Struct, Func, Gen. (2003)


Sequence Selection
For each sequence
For each sequence
CYANA
500 peptide structures
TINKER
500 peptide structures
AMBER Energy Evaluation
500 peptide structures
500 native peptide structures
Fold Specificity Calculation
RosettaAbinitio
1000 peptide structures
OREO
11 peptide structures
RosettaDock
110 complex structures
110 peptide structures
RosettaDesign
1 protein structure
• 22,000 peptide structures
• 22,000 complex structures
• 2000 protein structures
K* Calculation

Ensemble generation using RosettaDesign

- Algorithm Overview
  - Start from specified structure (110 starting structures for complex, 110 starting structures for peptide)
  - Change the **conformation** (rotamer) of a single amino acid
  - Accept or reject the change based upon the **Metropolis criterion**
  - Repeat
- Rosetta Design uses **backbone-dependant rotamer library** (≈150 rotamers for all amino acids at a given site) of Dunbrack and Cohen
- Generate **200 conformations** per starting structure (yields 200 × 110 = **22,000 total structures** each in the complex ensemble and the peptide ensemble)

Kuhlman B et al., Proc. Natl. Acad. Sci. (2000)


Sequence Selection
For each sequence
For each sequence
CYANA
500 peptide structures
TINKER
500 peptide structures
AMBER Energy Evaluation
500 peptide structures
500 native peptide structures
Fold Specificity Calculation
RosettaAbinitio
1000 peptide structures
OREO
11 peptide structures
RosettaDock
110 complex structures
110 peptide structures
RosettaDesign
1 protein structure
• 22,000 peptide structures
• 22,000 complex structures
• 2000 protein structures
K* Calculation

Designing an HIV-1 gp41 inhibitor

With Prof. R. Siliciano, Johns Hopkins Univ.

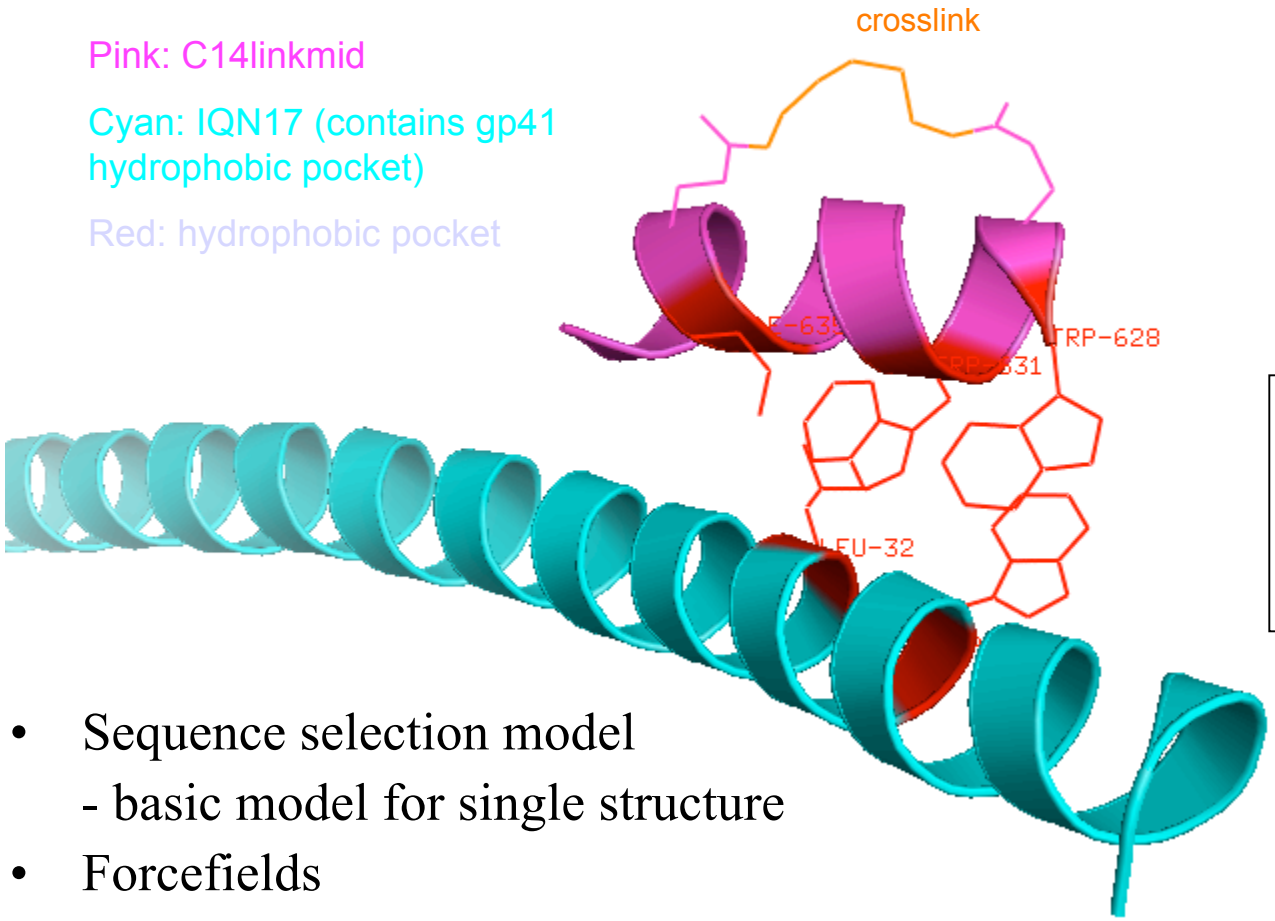
De novo design of HIV-1 gp41 inhibitor

- We used the C14linkmid-gp41 core complex crystal structure as design template

Pink: C14linkmid

Cyan: IQN17 (contains gp41 hydrophobic pocket)

Red: hydrophobic pocket



Crystal structure at resolution 1.9 Å by P.S. Kim group, 2002
PDB code: 1GZL

- Sequence selection model
  - basic model for single structure
- Forcefields
  - high resolution C^{α}-C^{α} forcefield
  - high resolution centroid-centroid forcefield

De novo design of HIV-1 gp41 inhibitor

- **Mutation set for sequence selection**

crosslink

Position	628	629	630	631	632	633	634	635	636	637	638	639
native residue	W	E	E	W	D	R	E	I	E	N	Y	T
allowed residues	H	H + C	P + C	H	P + C	P + C	P + C	H	P + C	P + C	H	P + C

H: hydrophobic P: polar C: cysteine

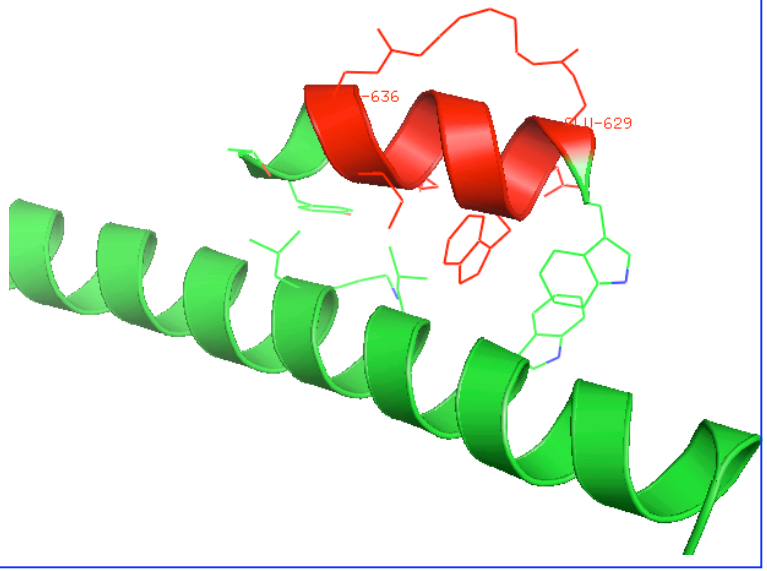
Problem complexity = 1.32×10^{12}

- **Biological constraints**

$$\sum_i y_i^{Arg} + \sum_i y_i^{Lys} - \sum_i y_i^{Asp} - \sum_i y_i^{Glu} = \text{native or native} \pm 1 \;\; \forall 629 \le i \le 636$$

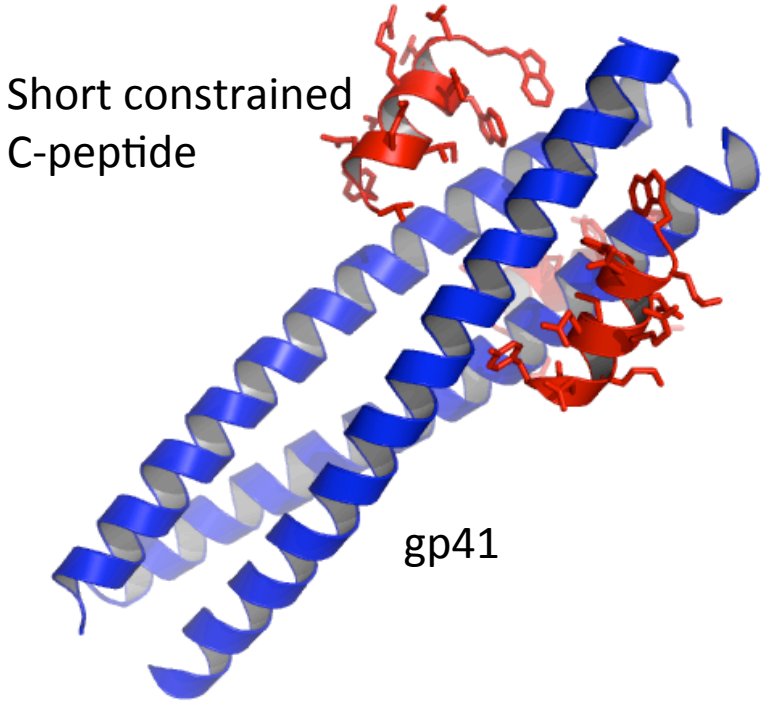
- **Number of mutations**
  - ≤3, ≤5, or no limit

- **Fold specificity stage**
  - C^{α}-C^{α} distance bounds: ±10% of those in crystal structure
  - dihedral angle bounds: ±10° around crystal structure



Theoretical prediction of HIV-1 gp41 inhibitors

Design Template



Sia, S.K., et al. PNAS (2002); PDB ID: 1gzl

Biological Constraints

Total charge on peptide is between 1 +/- native charge of peptide

Discovery

12-amino acid helical peptide

Mutation Set

Every residue allowed to mutate based upon Solvent Accessible Surface Area (SASA)

Complexity

4.46×10^{11} possible sequences

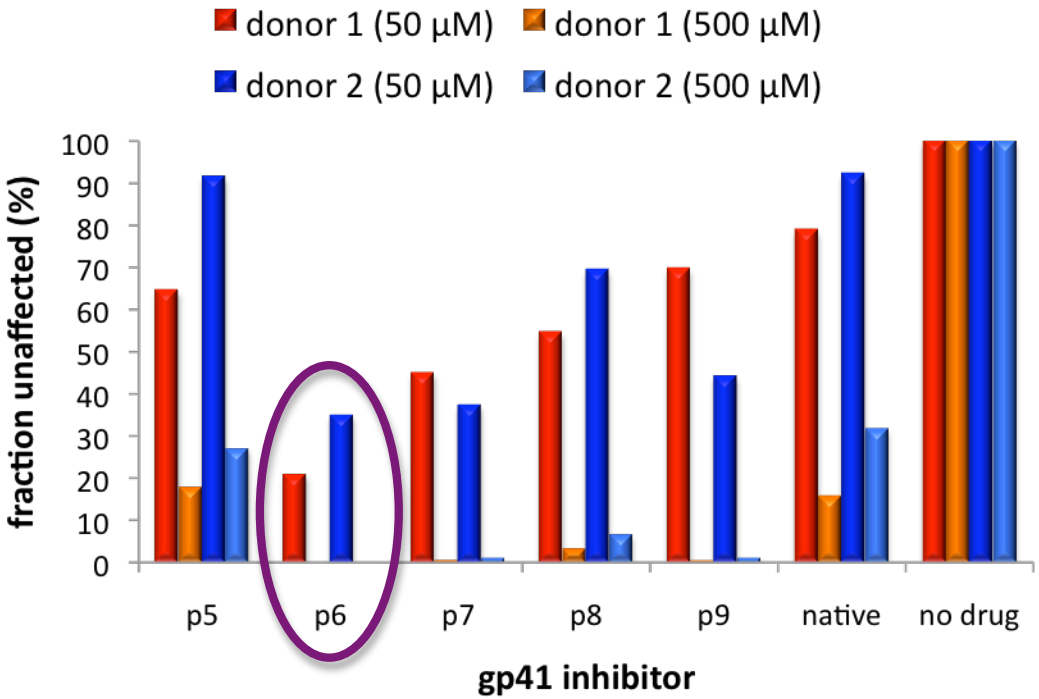
Selected Theoretical Results

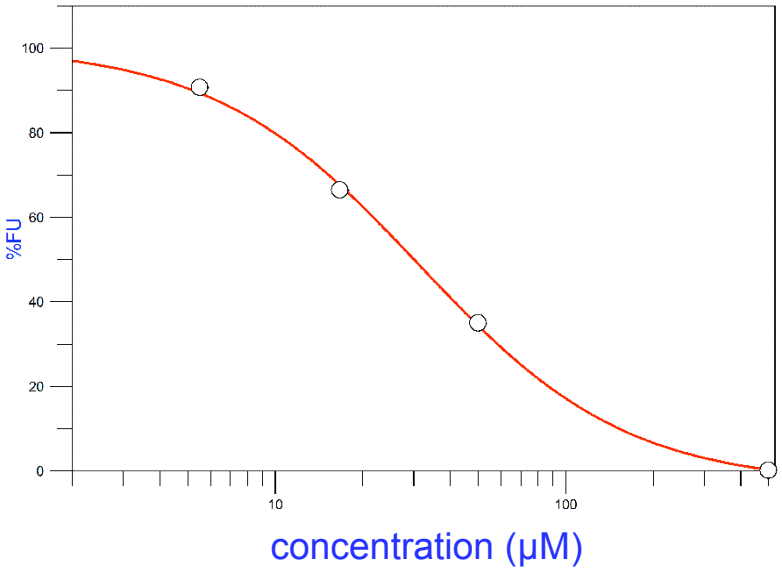
Stage Two Rank	Peptide Name	Approx. Binding Affinity (K*)
1	p5	$3.47\text{x}10^{+01}$
2	p6	$1.32\text{x}10^{+01}$
3	p7	$4.07\text{x}10^{+00}$
4	p8	$2.55\text{x}10^{+00}$
5	p9	$9.67\text{x}10^{-01}$
6	p10	$6.38\text{x}10^{-01}$
7	native	$2.42\text{x}10^{-01}$

Experimental results of HIV-1 gp41 inhibitors

Experimental Protocol

- normal (non-HIV-infected) person donates blood
- white blood cells are separated and all but the **CD4+ T cells** are removed
- T cells are activated with phytohemagluttinin for 3 days
- T cells are **infected** with a recombinant form of HIV that carries the gene for GFP (green florescence protein)
- cells that are infected**, fluoresce green**, cells that are not infected are not green
- infected cells & their sizes are detected by **flow cytometry**



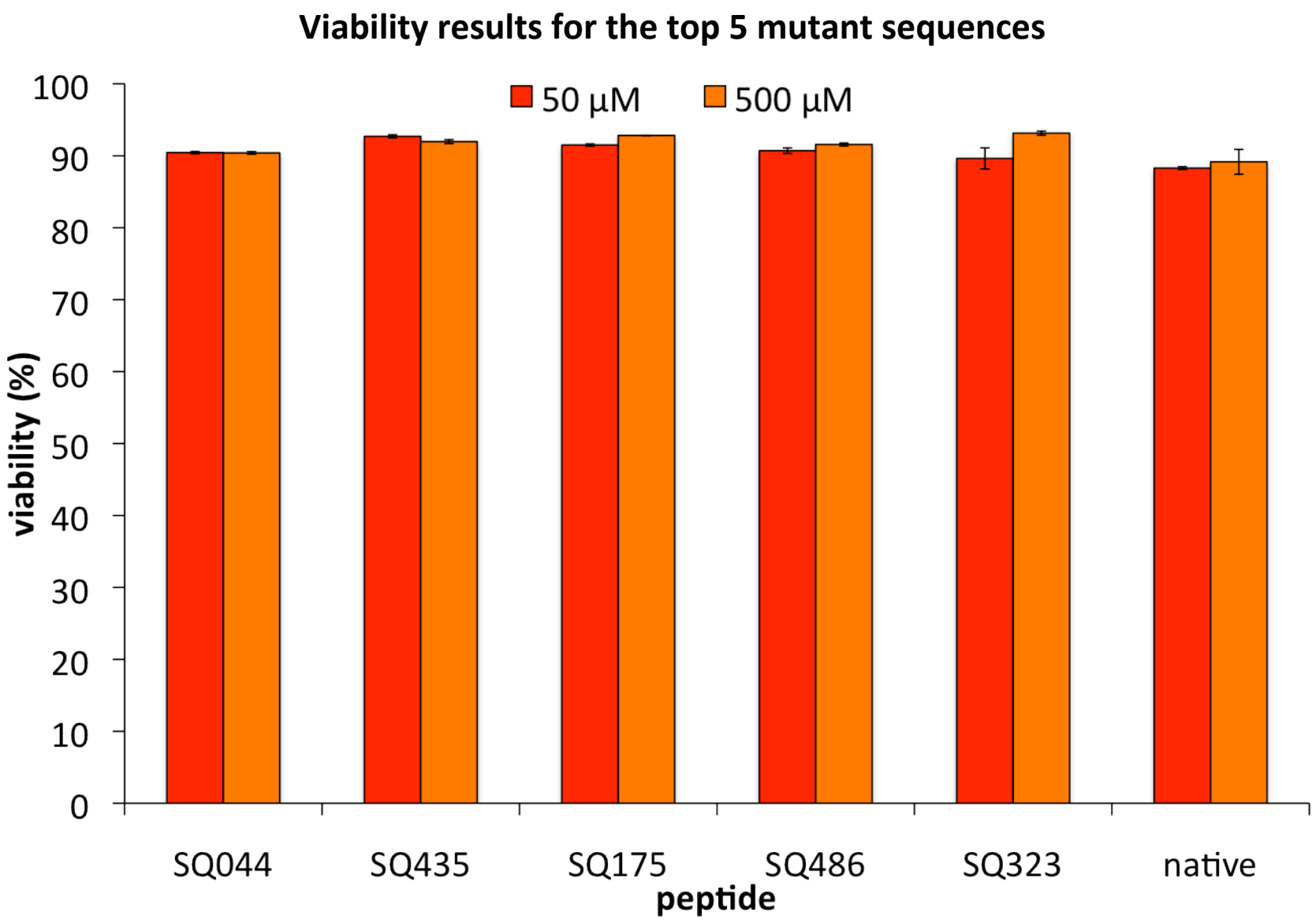


Parameter	Value	Std. Error
Y Range	103.7410	3.5366
IC 50	31.3415	2.4423
Slope factor	1.2159	0.1072
Background	-3.1846	2.6563

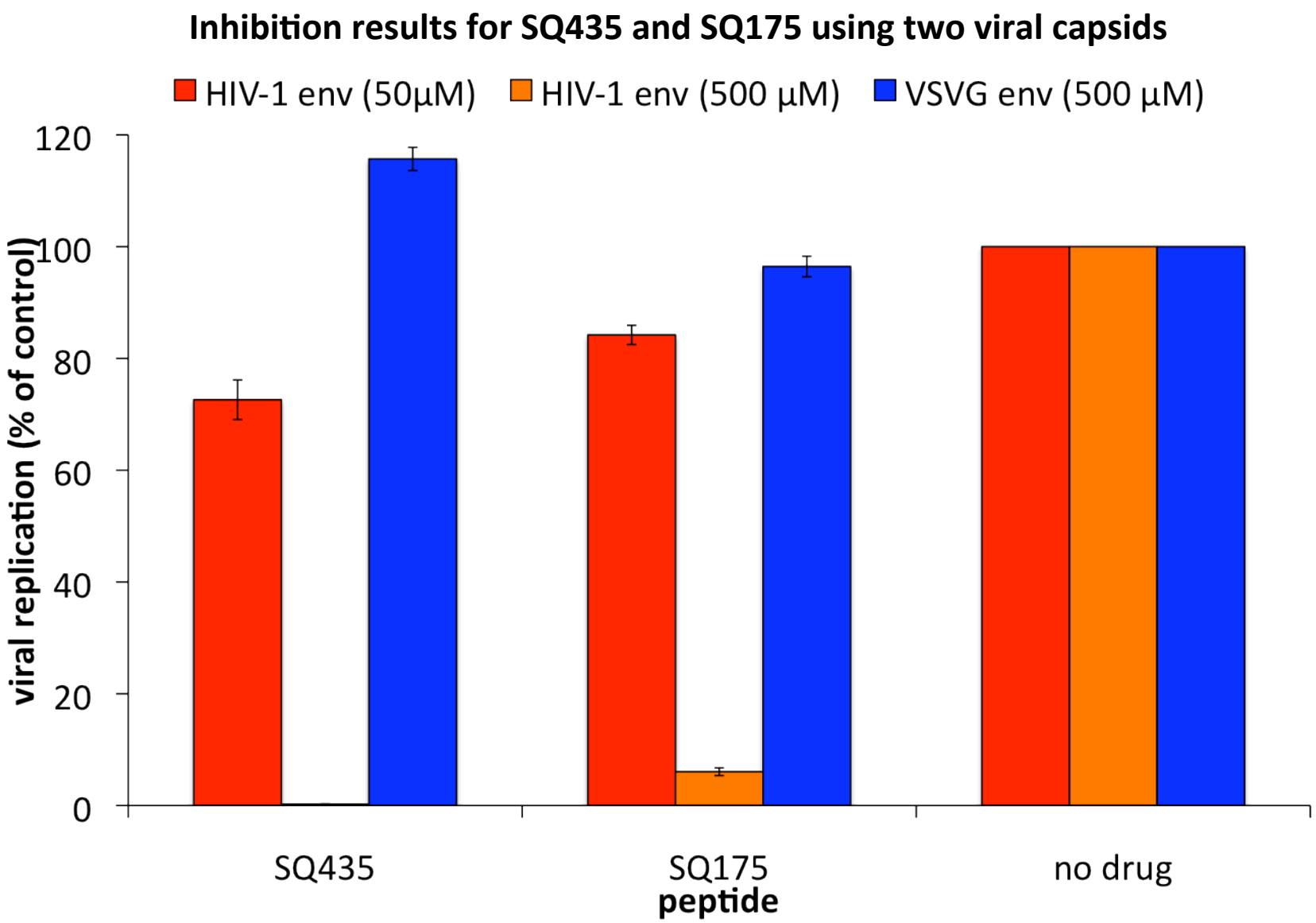
m=1.45

Shen et al., *Nature Medicine*, 2008

Are these peptides toxic?



Do the peptides target the HIV envelope?



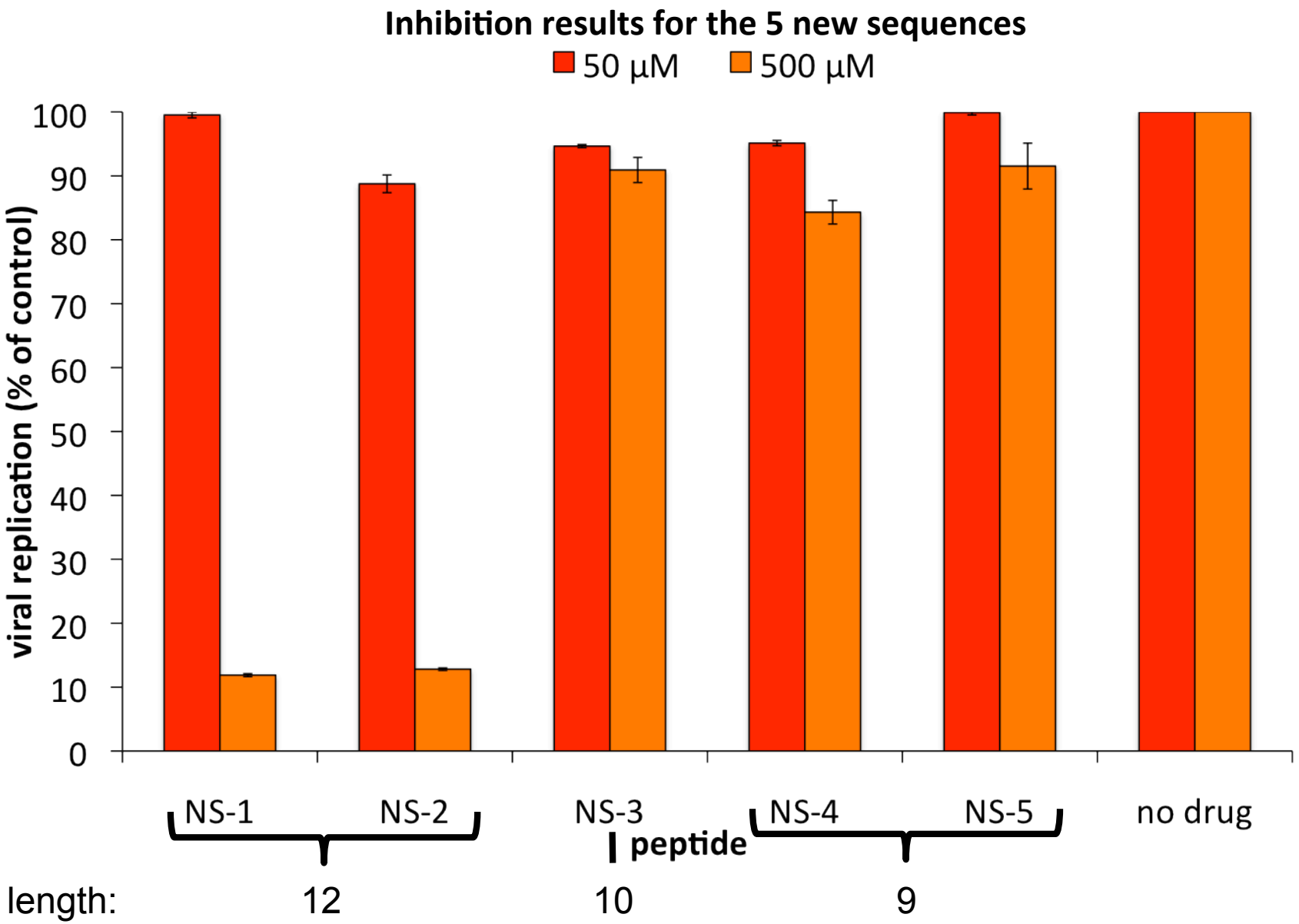
IC_{50} values for SQ435 and Enfuvirtide versus HIV-1 bearing five different envelopes

envelope	IC_{50} SQ435 (μM)			IC_{50} Enfuvirtide (nM)		
	donor 195	**donor 196**	***all data***	**donor 195**	**donor 196**	***all data***
HXB2	193	156	**174**	4.3	14.4	**7.5**
SF162	198	167	180	48	75	**63**
YU-2	148	73	**103**	37	49	**42**
C98v1e4	237	199	215	64	59	**62**
C109v1e4	195	257	227	17	39	**35**
all data	188 ± 11	158 ± 15	172 ± 10	23 ± 5	66 ± 11	40 ± 6

New sequences based upon the sequence of SQ435

Sequence Name	Length	Approx. Binding Affinity Rank	Sequence
SQ435	12	1	WCDWRDEWERYR
NS-1	12	2	WCDWRDEWERWR
NS-2	12	3	WCDWRDEWEWRR
native	12	4	WEEWDRWIENYT
NS-4	9	5	WCDWRDEWE
NS-3	10	6	WCDWRDEWER
NS-5	9	7	WCDWRDEWR

Can we reduce the length of the peptide?



Important Findings

- Six sequences were **ranked higher** than the native sequence based upon the approximate binding affinity calculations.
- Every mutant sequence except SQ044 showed **greater inhibition** than the native sequence.
- SQ435 is **the most potent inhibitor**, with an **IC_{50} of 29 – 253 μM** (3 – 15 fold more potent than native sequence) and **high donor-to-donor variability**.
- All peptides showed **minimal toxicity**.
- Peptides are HIV-1 **envelope specific**.
- **N43D mutation** causes **no change in susceptibility** to entry inhibition by SQ435. In contrast, mutation causes a **19-fold decrease in susceptibility of Enfuvirtide**.
- SQ435 is active against **multiple clinical isolates**.

Conclusions

- **New de novo protein design framework** for **flexible design templates**

- Novel sequence selection models for flexible design template with multiple structures
  - **framework allows for true protein backbone flexibility**

$$d^{L}_{C^{\alpha}-C^{\alpha}} \le d_{C^{\alpha}-C^{\alpha}} \le d^{U}_{C^{\alpha}-C^{\alpha}}$$
$$\phi^{L} \le \phi \le \phi^{U}$$
$$\varphi^{L} \le \varphi \le \varphi^{U}$$

- **Predictive framework is validated with experimental results**
  - Compstatin (13-residue): 16 & 45-fold improvement over native
  - C3a (77-residue): 15-fold improvement over native
  - HIV-1: gp41: on-going experimental validation

- Application of framework is now extended to other proteins
  (e.g., hβD-2, C3a, C5a, VCP, rhodopsin dimers, C3c, HIV-1 gp120)

Acknowledgments: NIH (R24,R01), NSF(ITR), EPA

Acknowledgements

Claire S. Adjiman	Imperial College	John L. Klepeis	D.E Shaw
Ioannis Akrotirianakis	SAS	Xiaoxia Lin	Harvard
Ioannis P. Androulakis	Rutgers University	Costas D. Maranas	Penn State
Meghan Bellows		Scott McAllister	ExxonMobil
Stavros Caratzoulas	University of Delaware	Clifford A. Meyer	DFCI, Harvard
William R. Esposito	Praxair	Rohit Rajgaria	
Ho Ki Fung		Heather D. Schafroth	Harvard
Zeynep H. Gumus	Cornell Medical School	Carl A. Schweiger	Pavillion Tech.
Steven T. Harding	Process Combinatorics	Karl M. Westerberg	CCSF
Marianthi G. Ierapetritou	Rutgers University		
Christopher Loose	MIT		
Brian Mickus	MIT	Michael Pieja	Yale University
Martin Taylor	Johns Hopkins		

Prof. John Lambris	University of Pennsylvania
Prof. Dimitris Morikis	University of California, Riverside
Prof. Panos Pardalos	University of Florida
Prof. Robert Siliciano	Johns Hopkins University
Prof. Arnold Neumaier	University of Vienna

National Science Foundation
National Institutes of Health
Enviromental Protection Agency