# Protein Folding, De Novo Protein Design, and Peptide Identification in Proteomics


Christodoulos A. Floudas

Princeton University

Department of Chemical Engineering

Program of Applied and Computational Mathematics

Department of Operations Research and Financial Engineering

Center for Quantitative Biology

# REVOLUTION OF GENOMICS

DNA (genome)

RNA (transcriptome)

Nucleus

Protein (proteome)

Sequence to Structure

Structure to Function

Protein - Protein

Nucleic Acid - Protein

Protein - Ligand

Feedback

Cell

Design of Drugs and Inhibitors

Metabolic Networks

Phenotype

# Outline

## From Sequence to Structure

- Structure Prediction in Protein Folding
  - ASTRO-FOLD
    - Helix Prediction
    - Beta Sheet Topology & Disulfide Bridges
    - 3-D Structure Prediction

## From Structure to Function

- De Novo Peptide Design
  - Sequence Selection
  - Fold Specificity
- Design of Inhibitors for Complement 3

## Peptide and Protein Identification via Tandem Mass Spectrometry

# Structure Prediction In Protein Folding: Outline

- Introduction to Protein Structure Prediction
- Free Energy Calculations in Oligo-peptides
- Prediction of Helical Segments
- Prediction of Beta Sheet Topologies
- Prediction of Loop Structures
- Derivation of Restraints
- Prediction of Protein Tertiary Structure

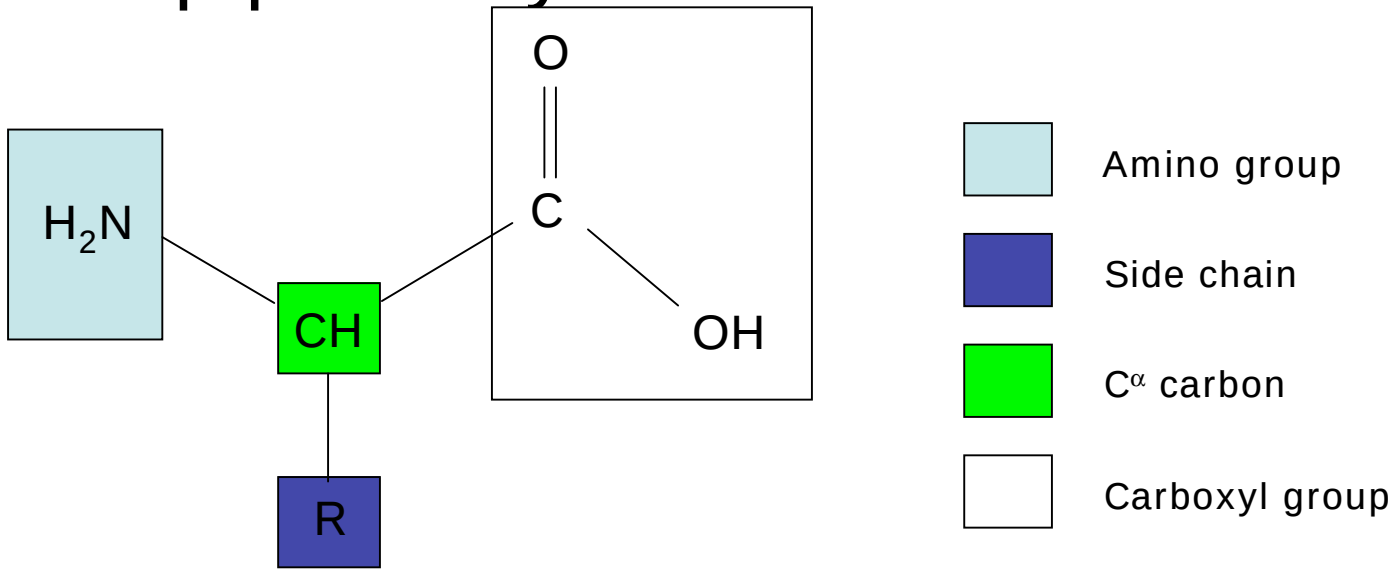
# Structure Prediction In Protein Folding

## Review Aricles

- Klepeis J.L., H.D. Schafroth, K.M. Westerberg, and C.A. Floudas, "Deterministic Global Optimization and Ab Initio Approaches for the Structure Prediction of Polypeptides, Dynamics of Protein Folding and Protein-Protein Interactions", Advances in Chemical Physics, 120, 265-457 (2002).
- Floudas C.A., "Research Challenges, Opportunities and Synergism in Systems Engineering and Computational Biology", AIChE Journal, 51, 1872-1884 (2005).
- Floudas C.A., H.K. Fung, S.R. McAllister, M. Monnigmann, and R. Rajgaria, "Advances in Protein Structure Prediction and De Novo Protein Design: A Review", Chemical Engineering Science, 61, 966-988 (2006).
- C.A. Floudas, "Computational Methods in Protein Structure Prediction", Biotechnology and Bioengineering, 97, 207-213 (2007).

# Protein Primary Structure

- Made up primarily of amino acids



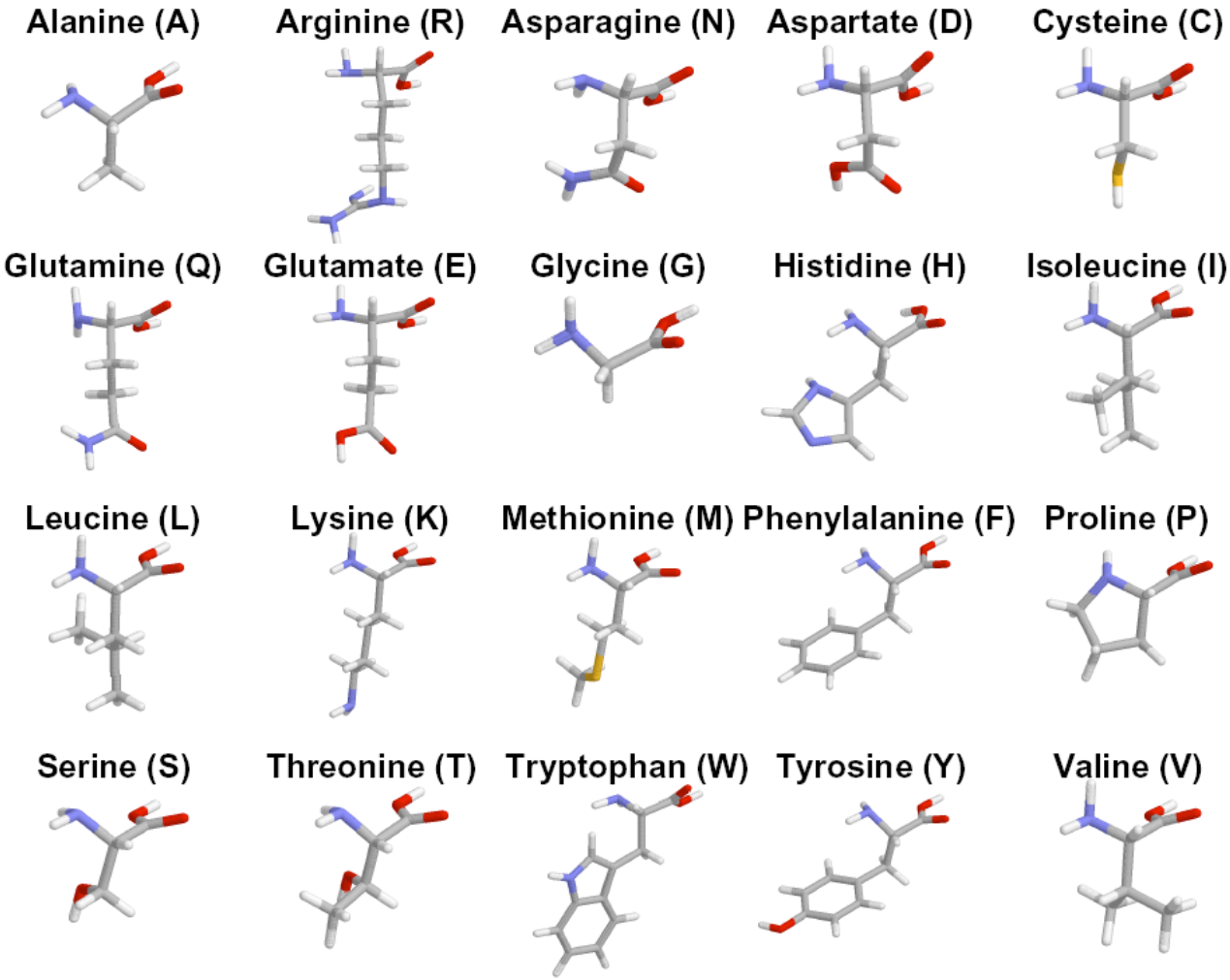

- This “alphabet” is often represented by a one letter abbreviation

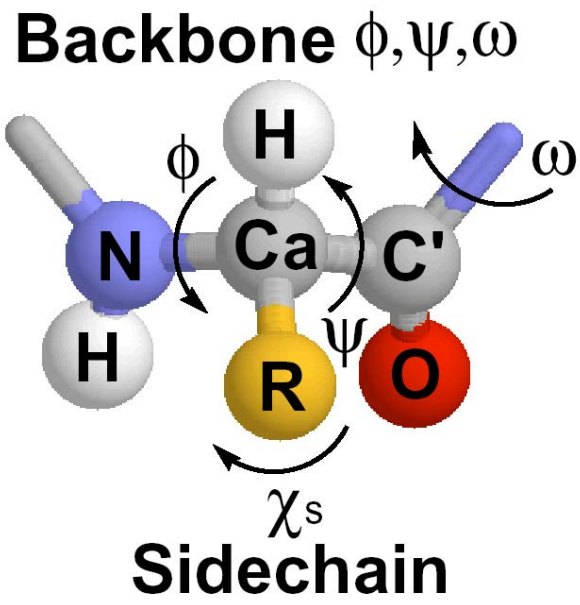
PDB: 1q4sA

**MHRTSNGSHATGGNLPDVASHYPVAYEQTLDGTVGFVIDEMTPERATASVEVTDTLRQRWGLVHGGAYCALAEMLA**
**TEATVAVVHEKGMMAVGQSNHTSFFRPVKEGHVRAEAVRIHAGSTTWFWDVSLRDDAGRLCAVSSMSIAVRPRRD**

# Proteins: Sequence of Amino Acids

- **20 letter alphabet**
- **200 amino acids/domain**
- **3M known sequences**
- **25K elucidated structures**

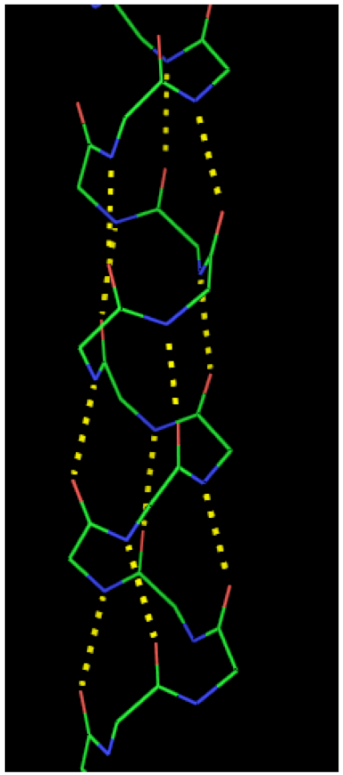


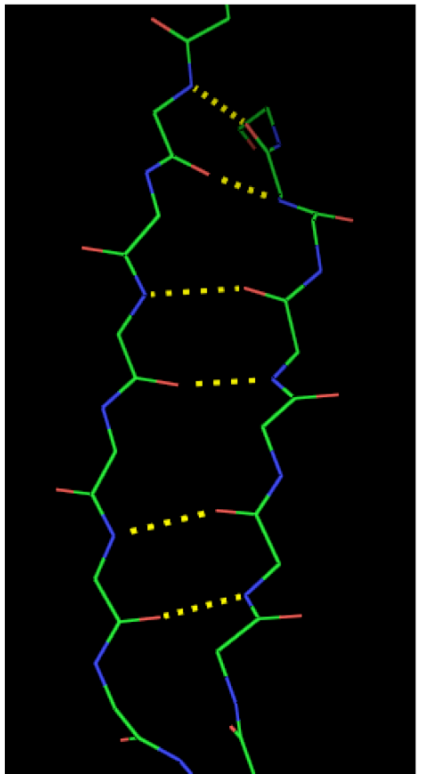
# Protein Secondary Structure

- Local structural motifs defined by hydrogen bonding patterns

α-helix

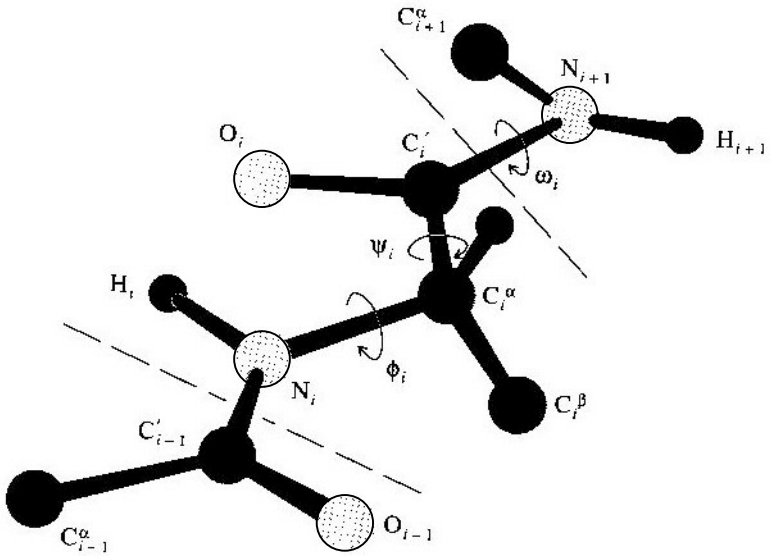


β-sheet



# Protein Dihedral Angles

- Fixed bond length
- Fixed bond angles
- Dihedral angles as variable representation



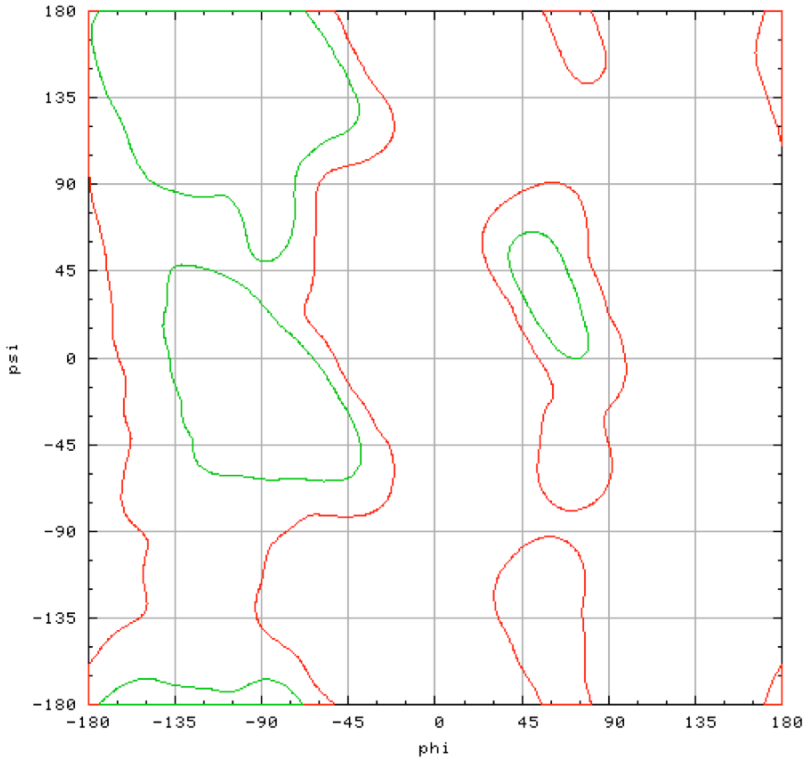
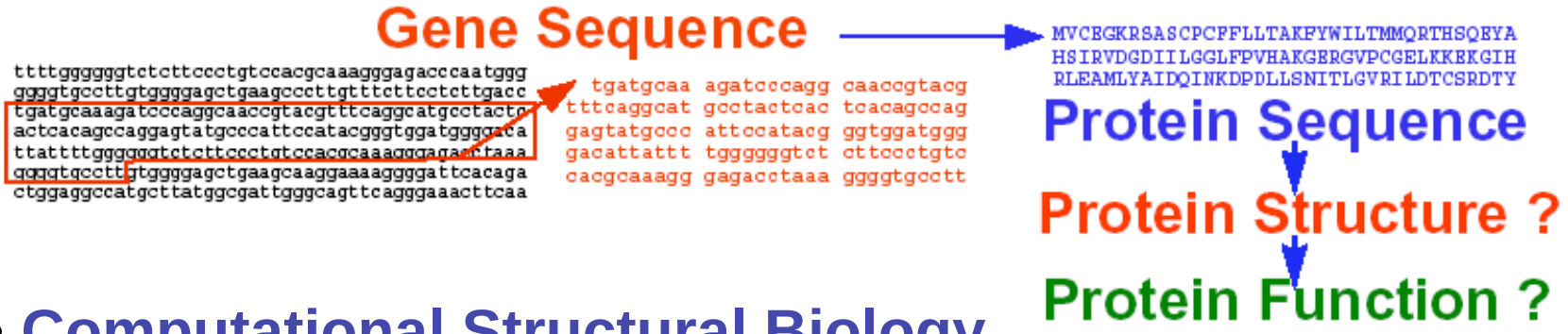




Image from Creighton, 1993

# Why Protein Folding ?

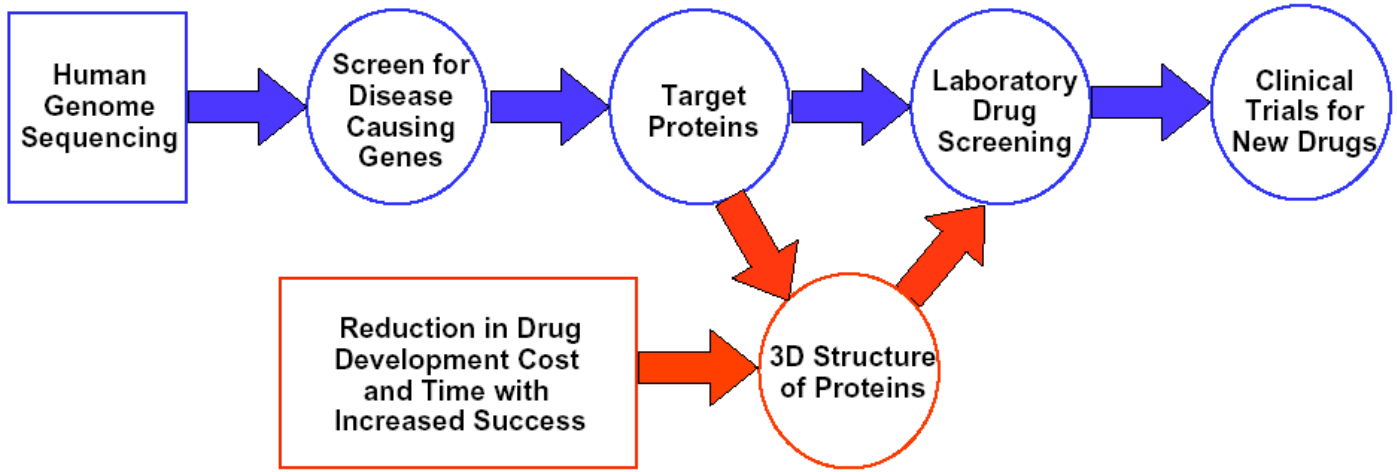
- **Human Genome Project** (Nature, Vol 409, 15 Feb 2001)
  - 3 x $10^9$ base pairs
  - ~ 3.1 x $10^4$ genes

→ $10^4$ functional proteins

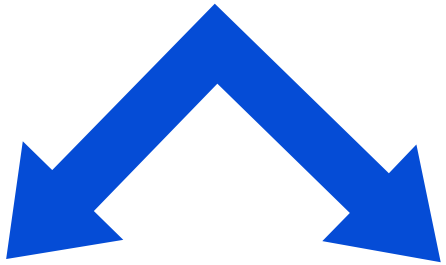



- **Computational Structural Biology**
  - Protein interactions
  - Role of mutations

→ Drug design



# Protein Folding Challenges



## Structure Prediction

Can we predict the 3-D structure from only the 1-D amino acid sequence ?

## Dynamics

Can we elucidate the mechanism of the folding process ?

How does the sequence of amino acids physically fold into the 3-D structure ?

# Protein Structure Prediction

## Amino acid sequence [PDB: 1q4sA ]

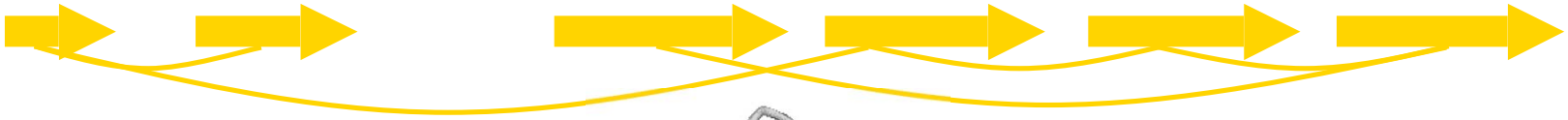
MHRTSNGSHATGGNLPDVASHYPVAYEQTLDGTVGFVIDEMTPERATASVEVTDTLRQRWGLVHGGAYCALAEMLA
TEATVAVVHEKGMMAVGQSNHTSFFRPVKEGHVRAEAVRIHAGSTTWFWDVSLRDDAGRLCAVSSMSIAVRPRRD

## Helical structure

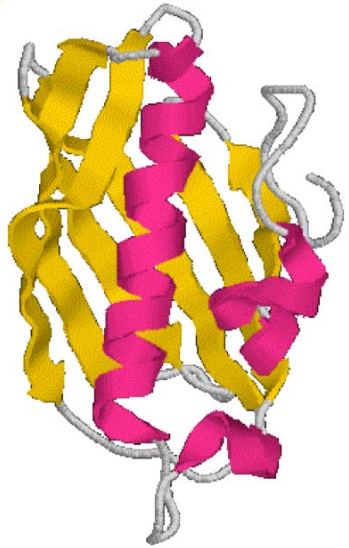
MHRTSNGSHATGGNLPDVASHYPVAYEQTLDGTVGFVIDEMTPERATASVEVTDTLRQRWGLVHGGAYCALAEMLA
TEATVAVVHEKGMMAVGQSNHTSFFRPVKEGHVRAEAVRIHAGSTTWFWDVSLRDDAGRLCAVSSMSIAVRPRRD

## Beta strand and sheet structure

MHRTSNGSHATGGNLPDVASHYPVAYEQTLDGTVGFVIDEMTPERATASVEVTDTLRQRWGLVHGGAYCALAEMLA
TEATVAVVHEKGMMAVGQSNHTSFFRPVKEGHVRAEAVRIHAGSTTWFWDVSLRDDAGRLCAVSSMSIAVRPRRD



## 3D Protein Structure



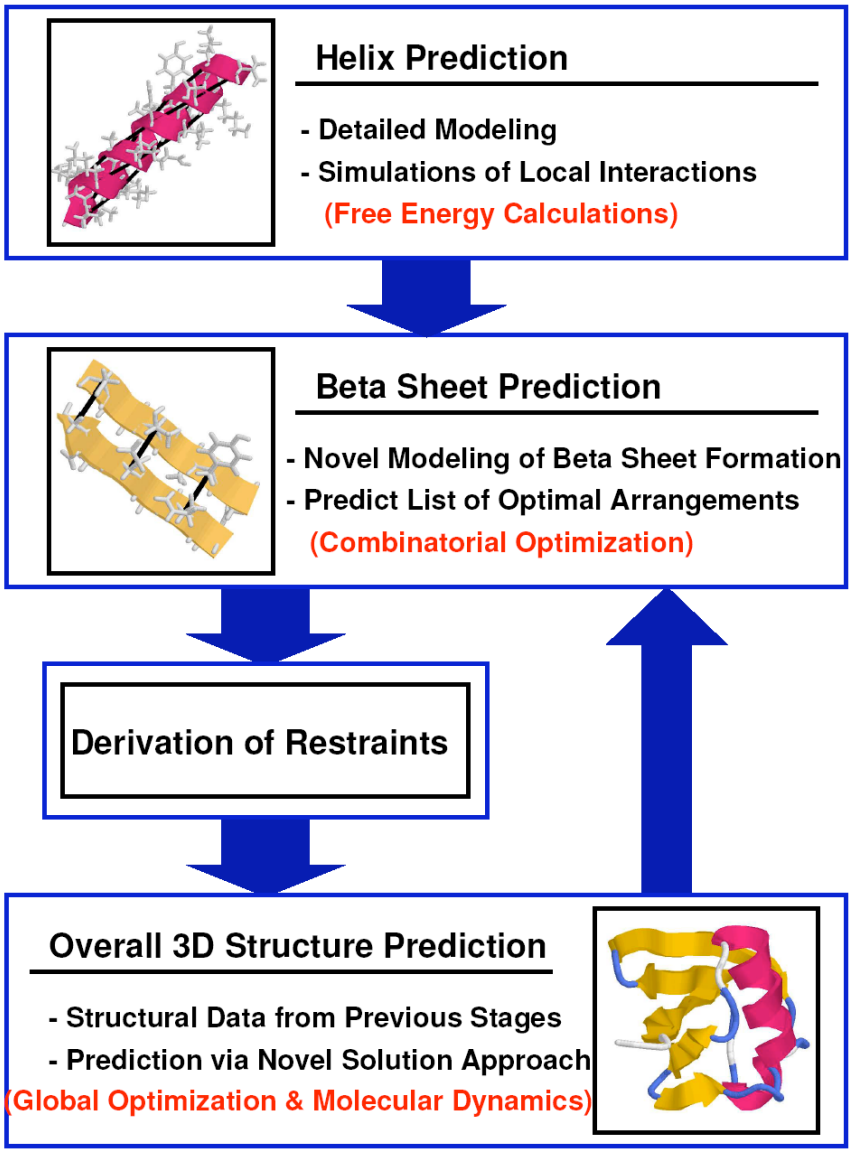
# Protein Folding: Advances

- **Homology Modeling / Comparative Modeling**
  - **The probe and template sequences are evolutionary related**
  - **Honig *et al*.; Sali *et al*.; Fischer *et al*.; Rost *et al*;**
- **Fold Recognition / Threading**
  - **For the query sequence, determine closest matching structure from a library of known folds by scoring function**
  - **Skolnick *et al*.; Jones *et al*.; Bryant *et al*.; Xu *et al*.; Elber *et al*.;**
  - **Baker *et al*.; Rychlewski & Ginalski; Honig *et al*.**

# Protein Folding: Advances

- **First Principles with Database Information**
  - **Secondary and/or tertiary information from databases/statistical methods**
  - **Levitt *et al*.; Baker *et al*.; Skolnick, Kolinski *et al*.; Friesner *et al*.**
- **First Principles without Database Information**
  - **Physiochemical models with most general application**
  - **Scheraga *et al*.; Rose *et al*.; Floudas *et al*.**

# ASTRO-FOLD




Klepeis, J.L and C.A. Floudas. *Biophys J* **85** (2003) 2119

# Structure Prediction In Protein Folding: Outline

- Introduction to Protein Structure Prediction
- Free Energy Calculations in Oligo-peptides
- Prediction of Helical Segments
- Prediction of Beta Sheet Topologies
- Prediction of Loop Structures
- Derivation of Restraints
- Prediction of Protein Tertiary Structure

# Free Energy Calculations in Oligopeptide Folding

**Relevant References:**


- **Maranas C.D., I.P. Androulakis and C.A. Floudas, "A Deterministic Global Optimization Approach for the Protein Folding Problem", DIMACS Series in Discrete Mathematics and Theoretical Computer Science, pp. 133-150, (1995).**
- **Androulakis I.P., C.D. Maranas and C.A. Floudas, "Prediction of Oligopeptide Conformations via Deterministic Global Optimization", Journal of Global Optimization, 11, pp. 1-34, June (1997).**
- **Klepeis J.L. and C.A. Floudas, "A Comparative Study of Global Minimum Energy Conformations of Hydrated Peptides", Journal of Computational Chemistry, 20, pp.636-654, (1999).**
- **Westerberg K.M. and C.A. Floudas, "Locating All Transition States and Studying Reaction Pathways of Potential Energy Surfaces", Journal of Chemical Physics, 110, pp.9259-9296, (1999).**
- **Klepeis J.L. and C.A. Floudas, "Free Energy Calculations for Peptides via Deterministic Global Optimization", Journal of Chemical Physics, 110, pp.7491-7512, (1999).**
- **Westerberg K.M. and C.A. Floudas, "Dynamics of Peptide Folding : Transition States and Reaction Pathways of Solvated and Unsolvated Tetra-Alanine", Journal of Global Optimization, 15, 261-297 (1999).**

# Goal

Develop a method for the theoretical prediction of native protein conformations via atomistic level modeling and global optimization

- Given : Only 1-D structural information (i.e., amino acid sequence)
- Select : Form of the atomistic level energy modeling which includes potential, solvation and entropic components
- Employ : Deterministic global optimization algorithm in order to locate the conformation exhibiting the global minimum energy

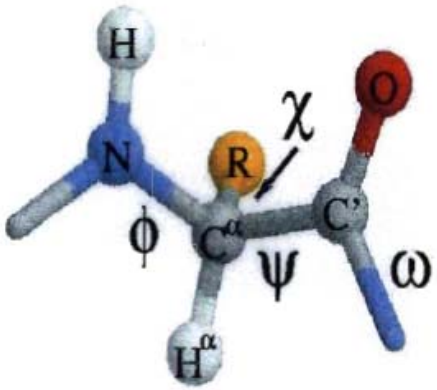
**Rationale**

Based on Anfinsen's thermodynamic hypothesis

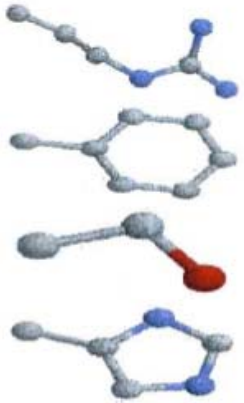
| | | |
|---|---|---|
| Biologically active (native) conformation | $\Longleftrightarrow$ | Global minimum energy conformation |

# Molecular Geometry

Peptides are linear polymers of the 20 naturally occurring amino acid residues

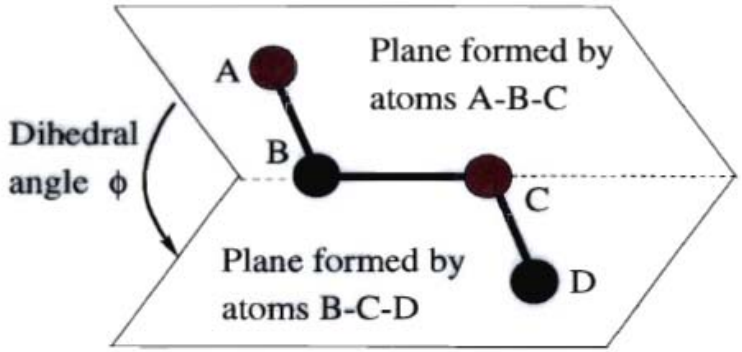



R =



etc.

Assume fixed bond distances and bond angles to transform Cartesian to internal coordinates ($\phi$, $\psi$, $\omega$, $\chi$)



# Potential Energy Modeling

Potential energy is calculated using ECEPP/3 force field (Némethy *et al.*, 1992)

Non-bonded

$$E_{pot} = \sum_{ij \in NB} \epsilon_{ij} \left[ \left( \frac{r^o_{ij}}{r_{ij}} \right)^{12} - \left( \frac{r^o_{ij}}{r_{ij}} \right)^{6} \right]$$

Hydrogen-bonded

$$+ \sum_{ij \in HB} \epsilon_{ij} \left[ \left( \frac{r^o_{ij}}{r_{ij}} \right)^{12} - \left( \frac{r^o_{ij}}{r_{ij}} \right)^{10} \right]$$

Electrostatic

$$+ \sum_{ij \in ES} \frac{332\ q_i q_j}{D r_{ij}}$$

Torsional

$$+ \sum_{k \in TOR} \frac{A_k}{2} (1 \pm \cos n_k \phi_k)$$

# Solvation Energy Modeling

Solvation energy is calculated implicitly

$$E_{sol} = \sum_{i=1}^{N} (S_i)(\delta_i)$$

- Neglect molecular nature of solvent (in contrast to explicit methods)
- Best suited for global searches
- Area-based and Volume-based methods

Solvent accessible volume shell model

- RRIGS (Augspurger *et al.*, 1996) volume shell calculations
- Avoid gradient discontinuities
- $\delta_i$ parameters based on experimental solubility data for organic molecules

# Free Energy Modeling

Entropic contributions **NOT** in most energy searches

Rigorous free energy modeling requires infinite sampling in order to associate accurate statistical weights with each microstate $\Longrightarrow$ NOT FEASIBLE

Boltzmann weight

$$\exp\left[\frac{-E(\theta_\gamma)}{k_B T}\right]$$

Harmonic Approximation

- Internal vibrational modes

$$\left(\frac{\partial^2 E}{\partial \theta^2}\right)_{\theta_\gamma}$$

Quasi-Harmonic Approximation

- Fluctuations mimic anharmonic trajectory
- Calculated directly from MC/MD simulations

$$(\theta - <\theta>)(\theta - <\theta>)^T$$

# Harmonic Approximation

- Consider partition function $Z$

$$Z = e^{-\frac{(E-TS)}{kT}} = e^{-\frac{E}{kT}} e^{\frac{S}{k}}$$

  - Boltzmann factor $\times$ Number of states available to system

- Harmonic approximation around stationary point

$$E(\theta) = E(\theta_\gamma) + \frac{1}{2}(\theta - \theta_\gamma) H_\gamma (\theta - \theta_\gamma)$$

  - Use $\nabla E(\theta_\gamma) = 0$

- $N_\theta$ independent harmonic oscillators with characteristic vibrational frequencies
- Each minima characterized by occupation of each normal mode
- Sum over energy states for harmonic oscillator

$$Z_\gamma = e^{-\frac{E_\gamma}{kT}} f(T) \prod_i^{N_\theta} \frac{1}{\lambda_i}$$

  - $\lambda_i$ represent eigenvalues of $H_\gamma$

$$\Downarrow$$

$$e^{\frac{S}{k}} \propto \prod_i \frac{1}{\lambda_i} \implies S \propto -k \ln\left[\mathrm{Det}(H_\gamma)\right]$$

# Free Energy Model

Harmonic entropy approximation for local energy minima $\longrightarrow$ FEASIBLE

- For each local minima ($\gamma$) calculate harmonic entropy $S_\gamma^{har}$

$$S_\gamma^{har} = -\frac{k_B}{2} \ln\left[\mathrm{Det}(\mathrm{H}_\gamma)\right]$$

- Harmonic entropy approximates shape of energy well using second derivative information [ $\mathrm{Det}(\mathrm{H}_\gamma)$]
  Wide minima give small values :
  Less negative harmonic entropies
  Narrow minima give large values :
  More negative harmonic entropies

- Calculate free energy at temperature $T$ using local energy minimum ($E_\gamma$) and $S_\gamma^{har}$

$$F_\gamma^{har} = E_\gamma + \frac{k_B\, T}{2} \ln\left[\mathrm{Det}(\mathrm{H}_\gamma)\right]$$

Requires an adequate ensemble of local minima to accurately represent statistical weights

# Generating Low Energy Ensembles

Rigorous $\alpha$BB Approaches

Formulation A

$$\begin{aligned} & \min_{\theta} && E(\theta) \\ \text{s.t.} \quad & (E^* - E) && + \epsilon^* < 0 \\ & \theta_i^L \leq && \theta_i \leq \theta_i^U \quad , \quad i = 1, \ldots, N_\theta \end{aligned}$$

- Constraint sets lower bound on global energy
- Increase $E^*$ to get next lowest minimum
- Iterate - full global optimization at each iteration

Formulation B

$$\frac{\partial E(\theta)}{\partial \theta_i} = 0, \quad i = 1, \ldots, N_\theta$$

- System of nonlinear equations
- Finds all stationary points
- $\alpha$BB reformulation

# $\alpha$BB Algorithm

1 Initialize best upper bound (BUB) to $+\infty$

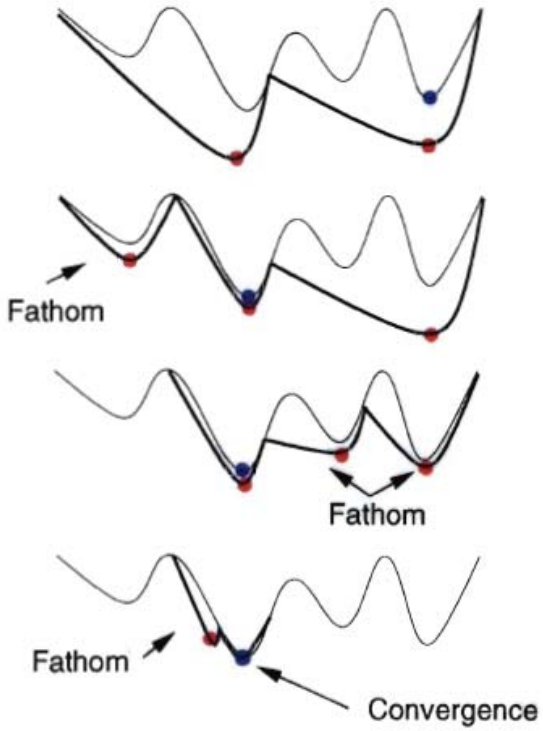
2 Partition domain along one dimension

3 Find lower bound in each subdomain

- Construct convex underestimators (L)
- Minimize and store each lower bound
- Fathom region if lower bound > BUB

4 Find an upper bound in each subdomain

- Locally minimize energy function
- Update BUB as min of all upper bounds




5 Select subdomain with lowest lower bound value for further partitioning

6 Terminate if BUB and lower bound within specified tolerance. Eise return to Step 2.

# Convex Underestimation

$$L(\mathbf{x}) = E(\mathbf{x}) + \sum_{i=1}^{N_{var}} \alpha_i (x_i^L - x_i)(x_i^U - x_i)$$

$$\alpha_i = \max\{0, -\frac{1}{2}\lambda_{min}\} \ , \ x_i \in [x_i^L, x_i^U]$$

## Properties

- $L$ is a valid underestimator of $E$
- $L$ matches $E$ at all corner points of the box constraints
- $L$ is convex in the current box constraints
- Maximum separation between $L$ and $E$ is bounded and proportional to $\alpha$ and to the square of the diagonal of the current box constraints (ensures $\epsilon$ tolerances)
- Underestimators $L$ constructed over supersets of the current set are always less tight than the underestimator constructed over the current box constraints for every point within the current box constraints

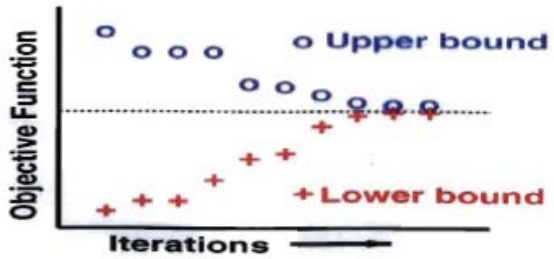
# $\alpha$BB Global Optimization

$$\min_{\theta} \quad E(\theta) \;=\; \min_{\theta} \quad \left[E_{pot}(\theta) \;+\; E_{sol}(\theta)\right]$$
$$\text{s.t.} \quad \theta_i^L \;\leq\; \theta_i \;\leq \theta_i^U \quad , \quad i = 1, \ldots, N_\theta$$

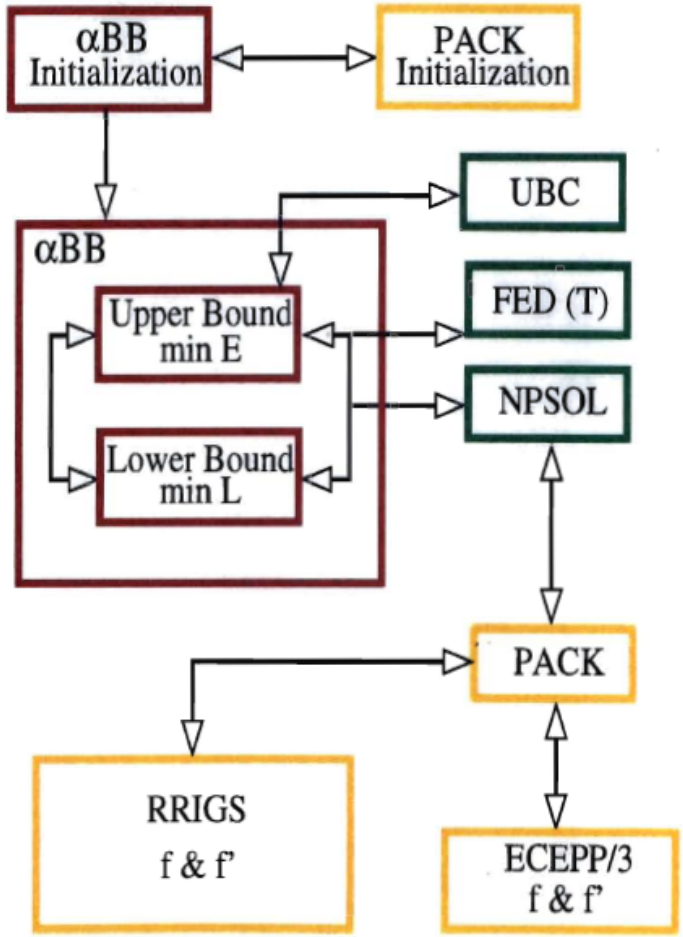
Conformational energy landscapes are highly nonconvex and require efficient global optimization methods independent of initial point selection

**$\alpha$BB** guarantees convergence to the global minimum

- Branch-and-bound framework
- Converge by solving a series of optimization problems which generate a non-increasing upper bound and a non-decreasing lower bound
- Upper bounds are local minima of original $\mathbf{E(x)}$
- Lower bounds are local minima of convex underestimating functions $\mathbf{L(x)}$



# Algorithmic Interface



# $\alpha$BB Based Free Energy Algorithm

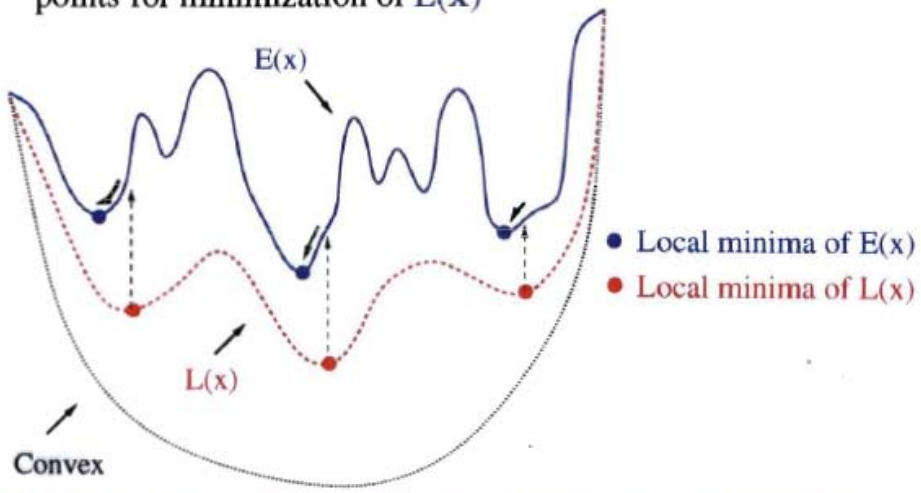
Harmonic free energy calculations require methods for finding low energy local minima

Two $\alpha$BB based approaches capitalize on :

- Ability to identify domains (rather than points) of low energy
- Information gained from lower bounding functions L(**x**)

**Approach 1.** ED-$\alpha$BB : Energy directed $\alpha$BB

- Parametric variation of $\alpha$ values
- Initialize and locally minimize lower bounding functions multiple times in each domain
- Unique lower bound minima serve as initial points for minimization of E(**x**)



# Free Energy Directed Search

**Rationale :**

- Enhance search for free energy local minima
- Incorporate harmonic entropy contributions

**Approach 2.** FED-$\alpha$BB : Free energy directed $\alpha$BB

- Add $-TS_{\gamma}^{har}$ at each local minima of the upper and lower bounding functions
- Rigorous implementation converges to global free energy minima
- Generate temperature dependent ensemble
- Thermodynamic temperature becomes an input parameter

# Computational Studies

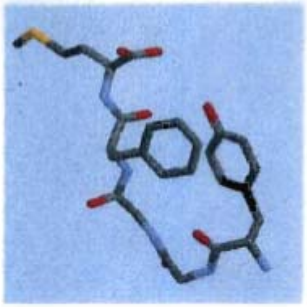
ED-$\alpha$BB and FED-$\alpha$BB applied to unsolvated and RRIGS solvated enkephalins

- 10 runs with initial $\alpha$ (and temperature) variation
- $\alpha$ reduction based on level in branch-and-bound tree
- 5 residues Tyr–Gly–Gly–Phe–(Met/Leu)
- 24 variables (dihedral angles)

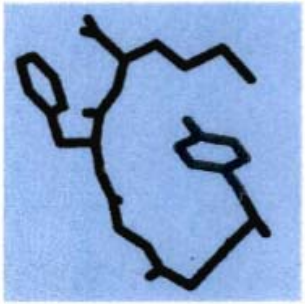
Initial Analysis

- Each run involved between 100,000 - 150,000 local minimizations of E($\mathbf{x}$)
- Unique structures identified using symmetry and requiring that at least one angle of any two structures differ by more than $50^o$
- Harmonic free energy of each structure calculated at several temperatures
- Rank and divide into bins (0.5 kcal/mol increments) above global minimum free energy
- Density of distinct metastable states follows a Boltzmann-like distribution within 5 kcal/mol of free energy global minimum

# Global Minimum Free Energy Structures

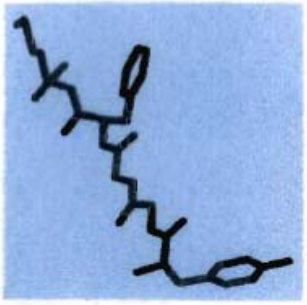


**Global Energy Minimum (T = 0 and $100K$)**

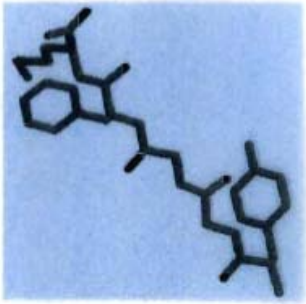


**A (T = $200K$)**

**B (T = $300K$)**



**C (T = $400K$)**



**D (T = $500K$)**

# Comparing Structures

Global minimum energy structure is generally **not** global minimum free energy structure

RRIGS solvated met-enkephalin

Table shows global minimum free energy and corresponding local minimum energy

| Temp → | 100 | 200 (**A**) | 300 (**B**) | 400 (**C**) | 500 (**D**) |
|---|---|---|---|---|---|
| $G_\gamma^{har}$ | -41.90 | -34.58 | -28.60 | -22.83 | -17.17 |
| $E_\gamma$ | -50.06 | -48.68 | -46.03 | -45.78 | -44.80 |

- Higher energy values offset by favorable entropic values at higher temperatures
- Global minimum energy and free energy structures are the same only at T = $100K$
- Global minimum free energy structures become more extended as temperature increases
- Better agreement with experimental structures when entropic effects are included

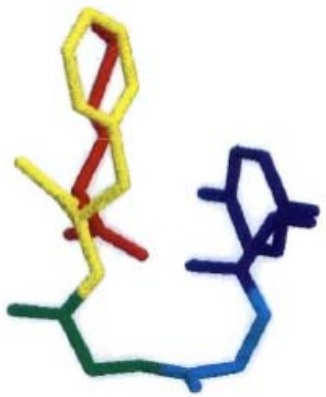
# Clustering Analysis

Cluster free energy minima based on structure

- Zimmerman codes for central residues
- Cumulative probability $F_{\text{cluster}} = -\frac{1}{\beta} \ln \sum p_i^{\text{approx}}$

| Temp | Code | Number | Prob | $F_{\text{cluster}}$ |
|---|---|---|---|---|
| 100 | DC*B | 107 | 0.532 | 0.125 |
| | C*DE | 990 | 0.232 | 0.291 |
| | CC*A | 1604 | 0.0636 | 0.547 |
| 300 | CD*A | 2128 | 0.263 | 0.796 |
| | C*DE | 1360 | 0.125 | 1.239 |
| | AAA | 327 | 0.111 | 1.309 |
| 500 | CD*A | 1966 | 0.0922 | 2.368 |
| | C*AE | 2088 | 0.0308 | 3.459 |
| | C*C*A | 1900 | 0.0279 | 3.555 |

GROUND STATE DC*B



CLUSTER @ 300K CD*A

# Transition States

Protein Folding is a dynamic transition between minima through a sequence of transition states

- First order saddle points are transition states
- Hessians have one negative eigenvalue

Alanine $\Longrightarrow$ $NH_2$-$CHCH_3$-C=0-OH

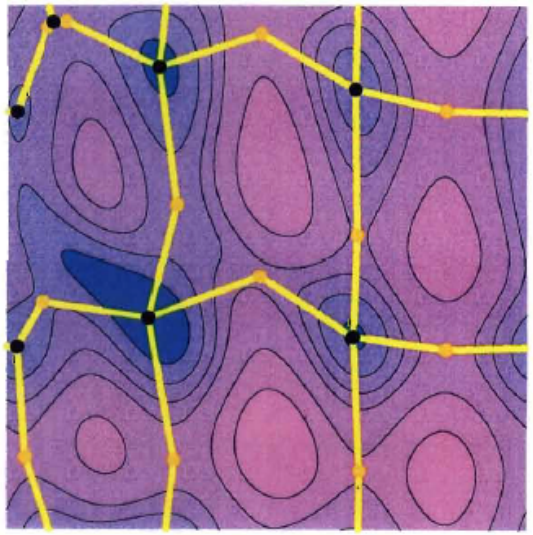
- $\omega$ and $\chi$ fixed at 180 degrees
- 7 Minima and 12 Transition States

How to find Transition States ?

$\Downarrow$

Follow low energy minima

| Follow | Min | TS |
|---|---|---|
| 3 | 7 | 9 |
| 4 | 7 | 11 |



# Eigenmode Following

Newton-Raphson method

- Traditional Newton-Raphson step
  Diagonalize Hessian and decompose gradient along eigenmodes

$$\Delta\theta = -H^{-1}g = -\sum_{i}^{N_\theta} \frac{g_i}{\lambda_i} e_i$$

- Minimize along positive eigenvalue modes
  Maximize along negative eigenvalue modes

Eigenmode Following (Tsai and Jordan, 1993)

- Shift eigenvalues by $L_i$

$$\Delta\theta = -\sum_{i}^{N_\theta} \frac{g_i}{\lambda_i - L_i} e_i$$

- Number of negative eigenvalues equals number of negative $(\lambda_i - L_i)$
- At each step follow eigenmode having largest overlap with previous eigenmode

# Pathway Connections

A. Union of

- 1000 lowest energy minima
- 1000 lowest free energy minima at 300 K

B. Locate first order transition states

- 2 possible initial steps (positive/negative)
- Follow each eigenmode (24 for met-enkephalin)

C. Identify Minimum-Transition-Minimum Triples

- 2 possible initial steps
- Follow each to minimum

| | # Total | # Min | # TS |
|---|---|---|---|
| Unsolvated | 51272 | 22775 | 28497 |
| Solvated | 76828 | 34722 | 42106 |

# Transition Rates

RRKM (Rice-Ramsperger-Kassel-Marcus) Theory

- Assume thermodynamic equilibrium at minima and transitions
- Harmonic approximation for partition functions

$$W_{\mathrm{min1}\to\mathrm{ts}\to\mathrm{min2}} = \frac{\prod_i^{N_\theta} f_{i,\mathrm{min1}}}{\prod_i^{N_\theta-1} f_{i,\mathrm{ts}}} \exp\left[\frac{-(E_{\mathrm{ts}} - E_{\mathrm{min1}})}{kT}\right]$$

$$W_{\mathrm{min2}\to\mathrm{ts}\to\mathrm{min1}} = \frac{\prod_i^{N_\theta} f_{i,\mathrm{min2}}}{\prod_i^{N_\theta-1} f_{i,\mathrm{ts}}} \exp\left[\frac{-(E_{\mathrm{ts}} - E_{\mathrm{min2}})}{kT}\right]$$

- Represents fraction of time in transition state

Applications

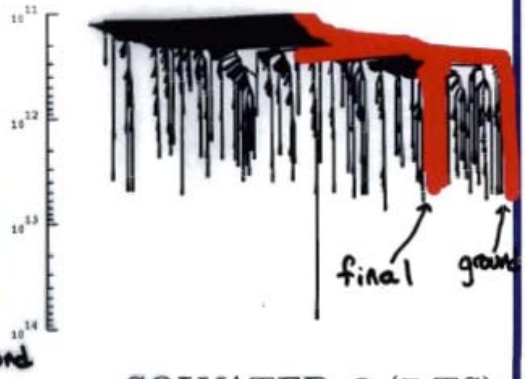
- Rate (Dis)Connectivity Graph (Becker and Karplus, 1997)
  - Use transition rate matrix to determine connectivity of minima
  - Identify slow (low frequency) and fast (high frequency) transitions
- Occupational probabilities
  - Solve Master equation
  - Identify time-dependent probabilities

# Tracing Pathways


UNSOLVATED (7 TS)

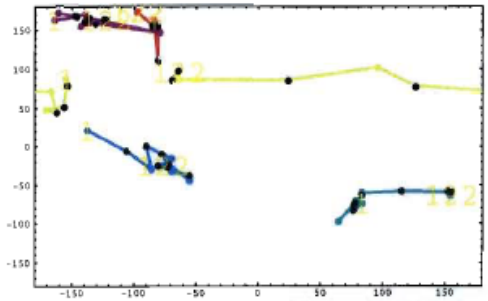

UNSOLVATED (9 TS)


SOLVATED (5 TS)

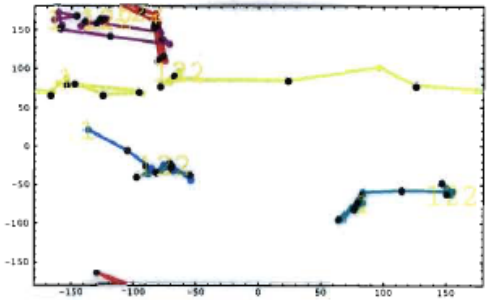



SOLVATED–2 (7 TS)

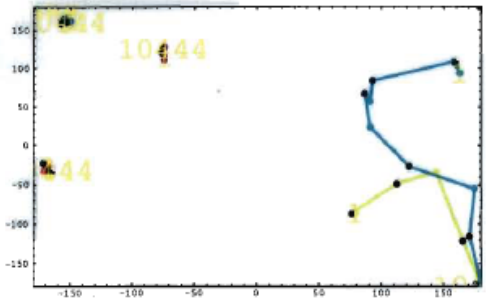
# Conformational Pathways



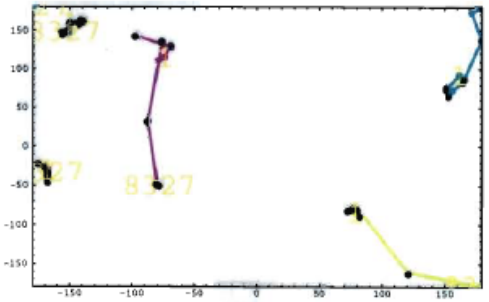
UNSOLVATED (7 TS)



UNSOLVATED (9 TS)



SOLVATED (5 TS)



SOLVATED–2 (7 TS)

# Structure Prediction In Protein Folding: Outline

- Introduction to Protein Structure Prediction
- Free Energy Calculations in Oligo-peptides
- Prediction of Helical Segments
- Prediction of Beta Sheet Topologies
- Prediction of Loop Structures
- Derivation of Restraints
- Prediction of Protein Tertiary Structure

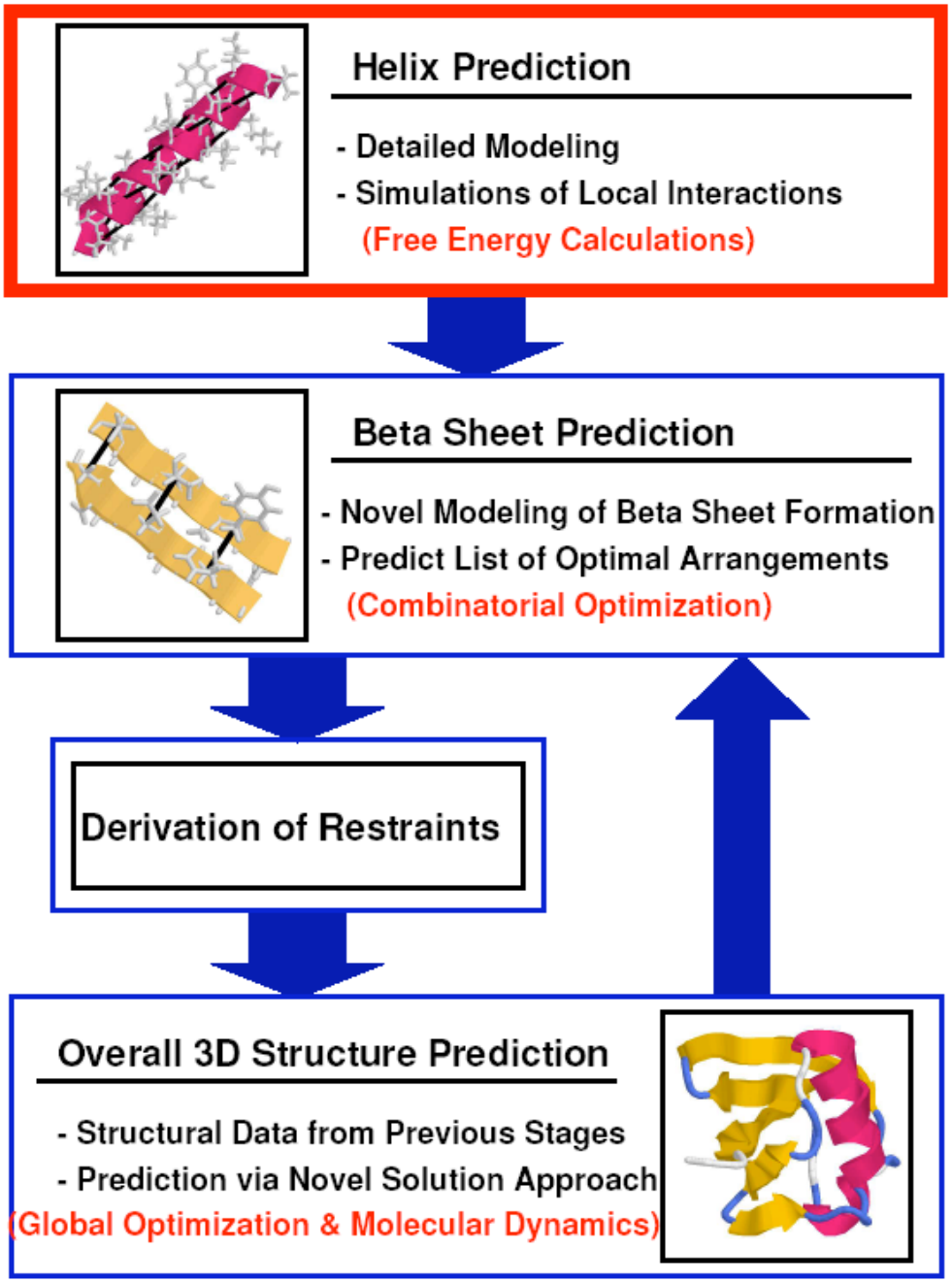
# Prediction of Helical Segments from First Principles

**Relevant References:**


- **Klepeis J.L. and C.A. Floudas, "Ab Initio Prediction of Helical Segments in Polypeptides", Journal of Computational Chemistry, 23, 245-266 (2002).**
- **Subramani A. and C.A. Floudas, “A Novel Approach for the Prediction of Helices and Beta Strands”, in preparation (2008).**

# ASTRO-FOLD

Klepeis & Floudas, 2002c



# Understanding Helix Formation

## Physical Characteristics of Helices

- Well defined backbones and hydrogen bonding patterns
- Different types of helices
  - $\alpha$ : hydrogen bonding every fourth residues (3.6)
  - $3_{10}$ : backbone turn every three residues

## Physical Understanding of Protein Folding

- Two competing explanations
  - Local forces : hierarchical folding
  - Non-local forces : hydrophobic collapse

## Experimental Evidence for Helix Formation

- Helix formation proceeds rapidly
- Sequence sufficient to identify initiation \ termination

Helix formation dominated by local forces

# Helix Prediction : Key Ideas

Klepeis & Floudas 2002a

## Overlapping oligopeptides

Decompose polypeptide to identify local sites of helix formation and termination

## Ensemble of low energy states

Calculate properties of proteins using data from many low energy states rather than a single state

## Free energy calculations Klepeis & Floudas 2000

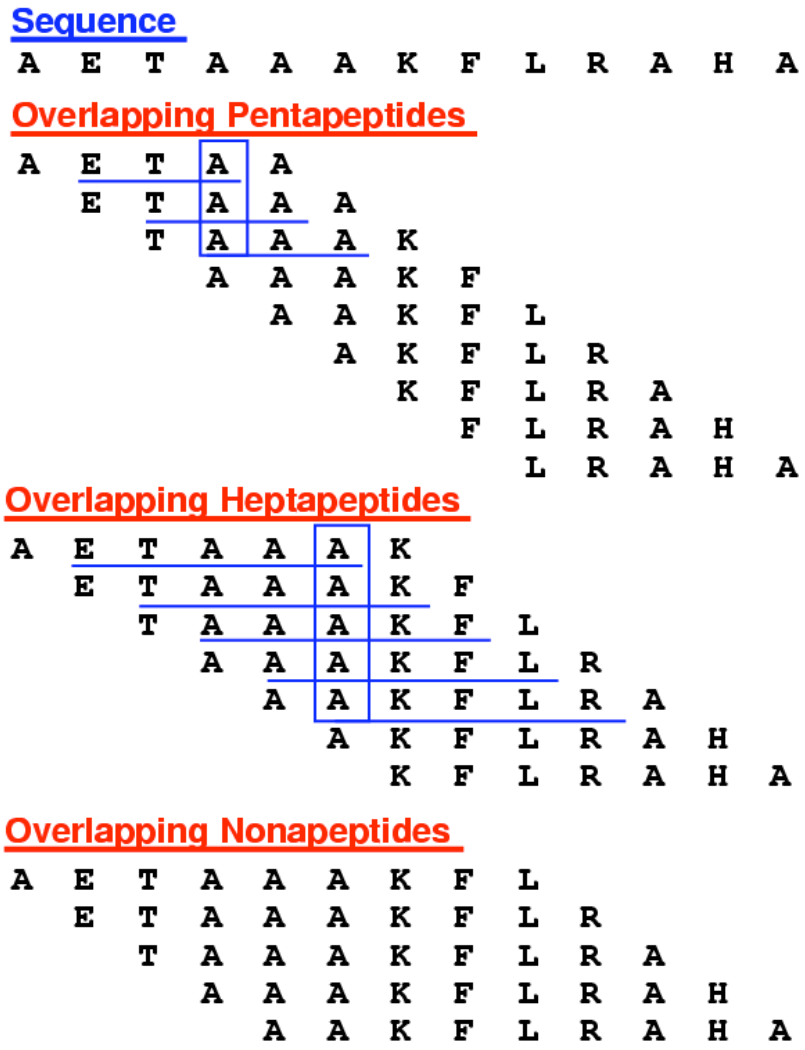
Model proteins using detailed energy calculations including entropic and solvation contributions

## Deterministic global optimization Floudas 2000

Predict low energy states using powerful global optimization approaches such as aBB

# Overlapping Oligopeptides

- Decompose polypeptide sequence into smaller oligopeptide sequences
  - Pentapeptides
  - Heptapeptides
  - Nonapeptides
- Capture local interactions governing helix formation
- Combine free energy calculations to get prediction



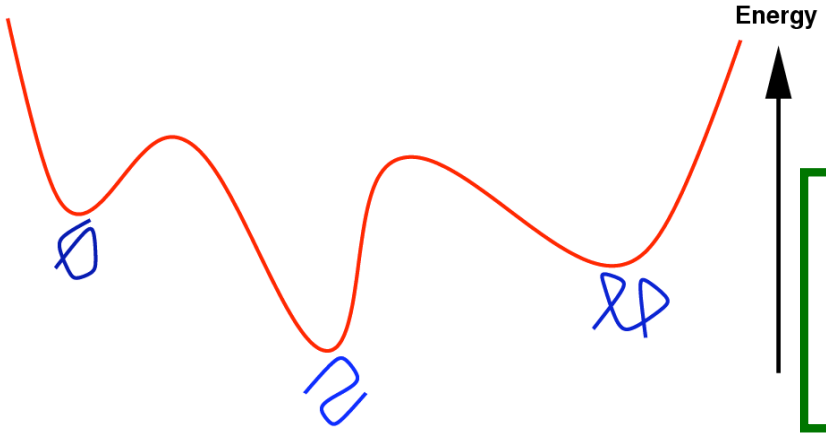
# Ensemble of Low Energy States

Klepeis & Floudas 1999


Generate low energy states along with global minimum energy state

## Mathematical formulation

- Nonconvex optimization problem
- Requires global optimization search




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

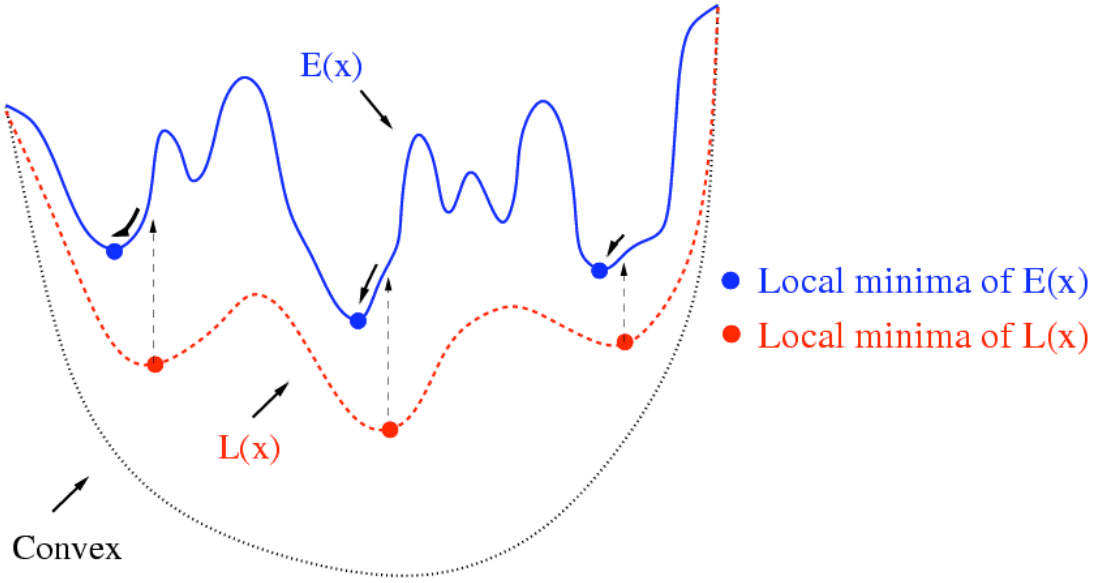
# Free Energy Algorithm

Klepeis & Floudas 1999

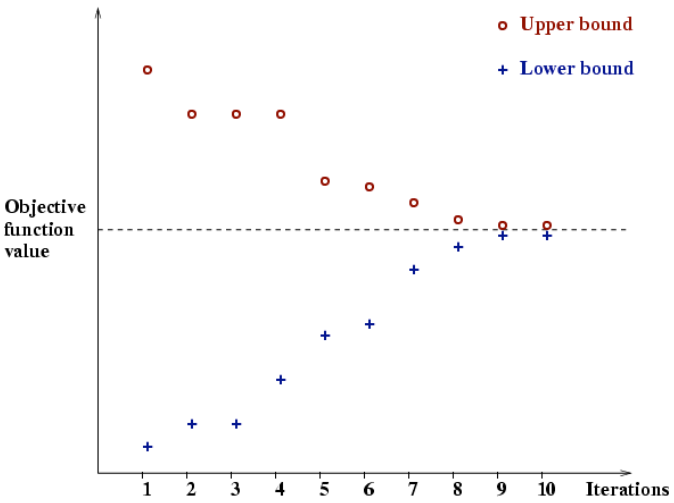
Free energy calculations require methods for finding low energy local minima

**αBB based approaches capitalize on**

- ability to identify domains of low energy
- information from lower bounding function L(x)
- initialize and locally minimize lower bounding functions multiple times in each domain
- unique lower bounding minima serve as initial points for minimizations of original E(x)



# Search Techniques




## αBB Deterministic Global Optimization

Floudas & coworkers 1994,1995,1996,1998,1999,2000

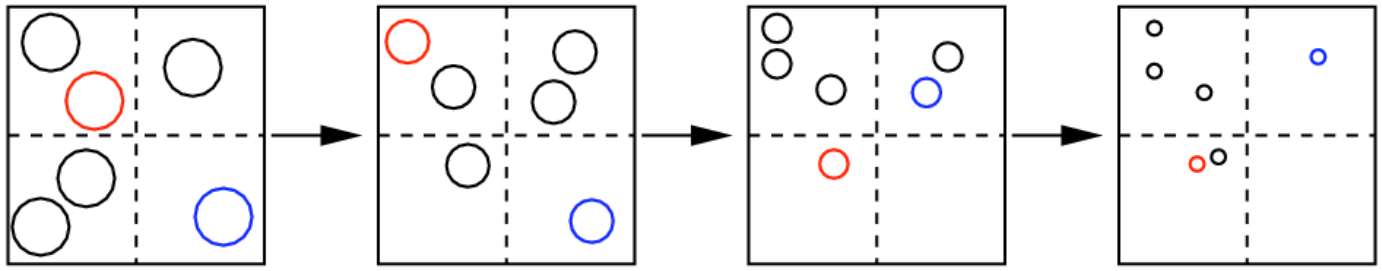
- based on branch-and-bound framework
- convergence through successive subdivision at each level in the b&b tree to generate non-increasing upper bounds (original problem) and non-decreasing lower bounds (convexified problem)
- guaranteed ε-convergence for $C^2$ NLPs

## Conformational Space Annealing (CSA)

Scheraga & coworkers 1997,1998

- stochastic prediction of global minimum energy state
- genetic algorithm updates produce low energy states
- anneal using deviation between energy states
- termination criteria is heuristic

**Decrease (anneal) size of cluster** →



# Free Energy Calculations

Klepeis & Floudas 2002a

$$F = F_{vac} + F_{cavity} + F_{solvation} + F_{ionization}$$

## Atomistic level free energy calculations

- include both enthalpic and entropic contributions
- model potential energy using semi-empirical force field
- employ harmonic approximation for entropic effects
- include cavity formation energy
- calculate solvation / ionization energies from solution of Poisson-Boltzmann equation Honig & coworkers 1988,1993,1995
- calculate total free energies for ensemble of low energy states
- employ efficient search techniques via global optimization

# Overall Free Energy

- **Potential**
  Scheraga & coworkers

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

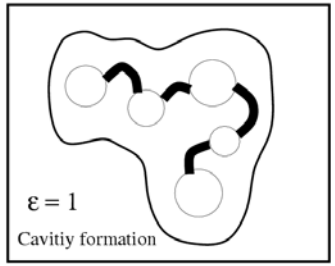
$F_{vac}$ -

- **Entropic**

$$-\frac{k_B}{2}\ln\left[\mathrm{Det}(\mathrm{H}_{\mathrm{vac},\gamma})\right]$$

$TS_{vac}$ +

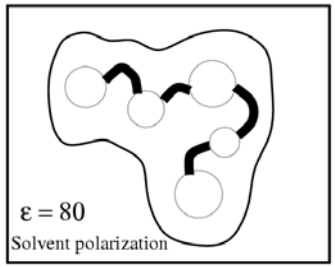
- **Cavity**
  Honig & coworkers
  1988, 1993, 1995




$F_{cavity} = \gamma(SA) + b$

$F_{cavity}$ +

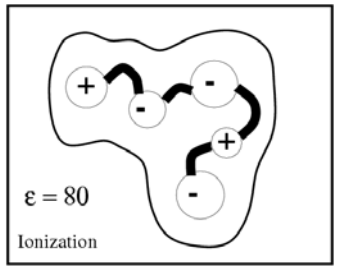
- **Polarization**
  Honig & coworkers
  1988, 1993, 1995




$F_{solv} = F_{polar}(\varepsilon=80) - F_{polar}(\varepsilon=1)$

$F_{solvation}$ +

- **Ionization**
  Honig & coworkers
  1988, 1993, 1995




$F_{ionize}(pH) = kT \ln(Z)$

$F_{ionization}$

# Atomistic Level Energy Modeling

$$F = F_{\text{vac}} + F_{\text{cavity}} + F_{\text{solvation}} + F_{\text{ionization}}$$

- Potential energy contribution to $F_{\text{vac}}$

$$F = E_{\text{vac}} - TS_{\text{vac}}$$

- Semi-empirical all atom force field

ECEPP/3 force field (Némethy *et al.*, 1992)

$$\begin{aligned} E_{\text{pot}} = & \text{ Non-bonded } \sum_{ij \in NB} \epsilon_{ij} \left[ \left( \frac{r_{ij}^o}{r_{ij}} \right)^{12} - \left( \frac{r_{ij}^o}{r_{ij}} \right)^{6} \right] \\ & + \text{ Hydrogen-bonded } \sum_{ij \in HB} \epsilon_{ij} \left[ \left( \frac{r_{ij}^o}{r_{ij}} \right)^{12} - \left( \frac{r_{ij}^o}{r_{ij}} \right)^{10} \right] \\ & + \text{ Electrostatic } \sum_{ij \in ES} \frac{332\, q_i q_j}{D r_{ij}} \\ & + \text{ Torsional } \sum_{k \in TOR} \frac{A_k}{2} (1 \pm \cos n_k \phi_k) \end{aligned}$$

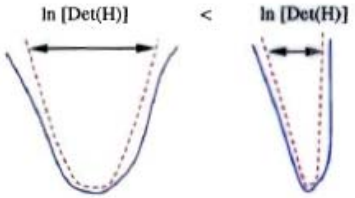
# Entropic Contributions

$$F = F_{\text{vac}} + F_{\text{cavity}} + F_{\text{solvation}} + F_{\text{ionization}}$$

- Entropic contribution to $F_{\text{vac}}$

$$F = E_{\text{vac}} - TS_{\text{vac}}$$

- Harmonic entropy approximation
- Mathematical approximation to entropy of each energy state




- For each unique energy state ($\gamma$) calculate harmonic entropy

$$S_{\text{vac},\gamma} = -\frac{k_B}{2} \ln\left[\text{Det}(\text{H}_{\text{vac},\gamma})\right]$$

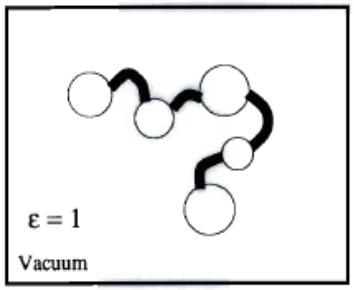
- Calculate free energy at temperature $T$ using local energy minimum ($E_{\text{vac},\gamma}$) and $S_{vac,\gamma}$

$$F_{\text{vac},\gamma} = E_{\text{vac},\gamma} + \frac{k_B\, T}{2} \ln\left[\text{Det}(\text{H}_{\text{vac},\gamma})\right]$$

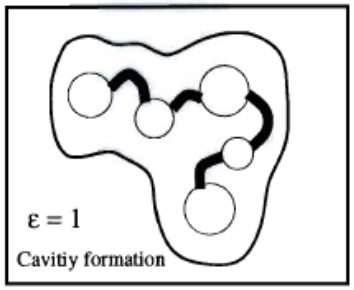
# Solvation Effects : Cavity Formation

$$F = F_{\text{vac}} + F_{\text{cavity}} + F_{\text{solvation}} + F_{\text{ionization}}$$

- Model cavity formation in aqueous environment
- Solvent accessible surface area correlation
- Macroscopic term based on empirical fit
  ($\gamma = 0.005$ kcal/mol $\AA$, $b = 0.860$ kcal/mol)




**Initial Conformer Free energy based on vacuum potential**



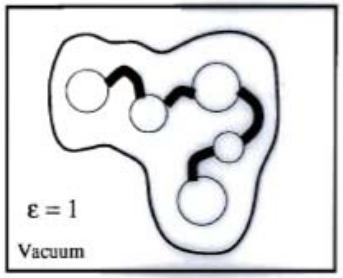

**$F_{\text{cavity}} = \gamma(SA) + b$**

(Honig & coworkers)

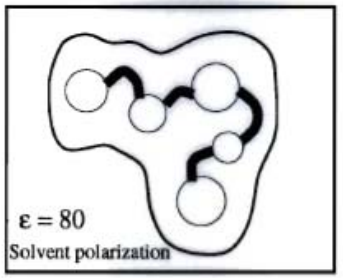
# Solvation Effects : Polarization

$$F = F_{\text{vac}} + F_{\text{cavity}} + F_{\text{solvation}} + F_{\text{ionization}}$$

- Calculate polarization effects in aqueous phase
- Difference of polarization free energies between vacuum and water environments
- Solve Poisson-Boltzmann equation for two systems where difference is dielectric constant




**Initial Conformer Free energy based on vacuum potential plus Cavity formation**




$$F_{\text{solv}} = F_{\text{polar}}(\varepsilon=80) - F_{\text{polar}}(\varepsilon=1)$$

(Honig & coworkers)

# Poisson-Boltzmann Calculations

## Nonlinear Poisson Boltzmann Equation

$$\nabla \cdot [\epsilon(\mathbf{r}) \nabla \cdot \phi(\mathbf{r})] - \kappa(\mathbf{r})^2 \sinh [\phi(\mathbf{r})] + \frac{4\pi \rho^f(\mathbf{r})}{kT} = 0$$

- $\epsilon(\mathbf{r})$ is the dielectric at position $r$
- $\phi(\mathbf{r})$ is the total electrostatic potential
- $\rho^f$ is the ionic charge density
- Solve by finite difference, boundary element or other numerical methods to get description of effective potential

Solution of Nonlinear Poisson Boltzmann Equation :

- Polarization Free Energy
- Ionization Free Energy

# Poisson-Boltzmann Calculations

## Polarization Free Energy

$$F_{\text{solv}} = F_{\text{polar}}(\epsilon = 80) - F_{\text{polar}}(\epsilon = 1)$$

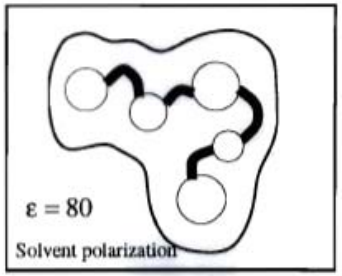
- Calculate **Reaction Field Energy** for two states : $\epsilon = 80$ and $\epsilon = 1$
- Calculate induced surface charge at the surface and then sum the potential at every charge

$$F_{\text{polar}} = \frac{1}{2} \sum_i \sum_s \frac{q_i \sigma_s}{|\mathbf{r_i} - \mathbf{r_s}|}$$

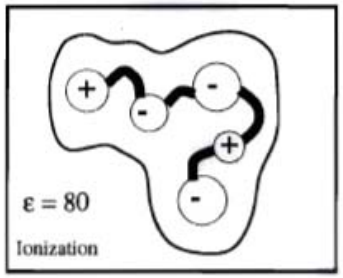
# Solvation Effects : Ionization

$$F = F_{\text{vac}} + F_{\text{cavity}} + F_{\text{solvation}} + F_{\text{ionization}}$$

- Additional calculations for ionization of titratable residues
- Thermodynamic cycle involves difference in free energy between neutral and protonated forms
- Decomposition into set of Poisson-Boltzmann calculations




**Free energy based on solvated system plus Cavity formation**




$$E_{\text{ionize}}(\text{pH}) = kT \ln(Z)$$

(Honig & coworkers)

# Poisson-Boltzmann Calculations

## Ionization Free Energy

- Consider partition function for all ionization states

$$F_{\text{ionize}}(\text{pH}) = kT \ln Z$$

$$Z = \sum_{i=1}^{2^N} \exp\left[-\Delta G_i / kT\right]$$

- Free Energy of $i$th state (consider all combinations)

$$\Delta G_i = \sum_{j=1}^{N} \left( x_j 2.303 kT \left(\text{pH} - \text{pK}_j\right) + \delta_j \sum_{1 \le k < j} \delta_k \Delta G_{jk} \right)$$

- Calculate intrinsic $\text{pK}_\text{a}$ : Difference between ionization of protein environment and of isolated aqueous phase

$$\text{pK}_j = \text{pK}_j^o - \gamma_j \Delta\Delta G_j / 2.303 kT$$

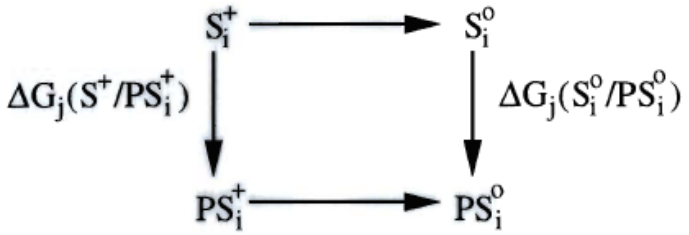
$$\Delta\text{pK}_j = \frac{\Delta\Delta G_j}{\gamma_j 2.303 kT}$$

# Poisson-Boltzmann Calculations

## Ionization Free Energy

- Obtain $\Delta\Delta G_j$ from Thermodynamic cycle

$$\frac{\Delta\Delta G_j}{\gamma_j} = \left(\Delta G_j(\mathrm{PS}_i^+/\mathrm{S}_i^+) - \Delta G_j(\mathrm{PS}_i^o/\mathrm{S}_i^o)\right)$$




- $\Delta G_j(\mathrm{PS}_i^+/\mathrm{S}_i^+)$ represents the change in free energy when moving the (ionized) ionizable group from an isolated aqueous environment into the protein environment.

- $\Delta G_j(\mathrm{PS}_i^o/\mathrm{S}_i^o)$ represents the same transition but for the neutral form of the ionizable group.

- Individual $\Delta G_j$ terms can be further decomposed into reaction field effects and permanent dipole effects

$$\Delta G_j = \Delta G_j^{\text{rxn field}} + \Delta G_j^{\text{perm dipole}}$$

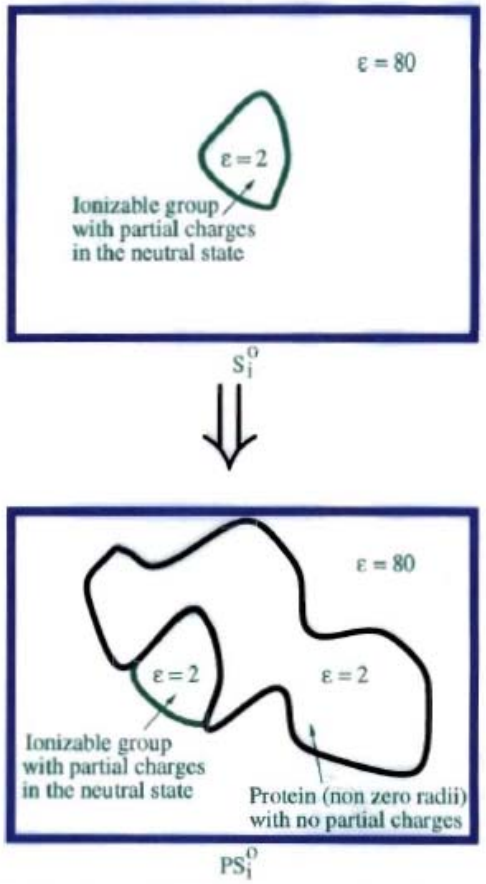
# Poisson-Boltzmann Calculations

## Reaction Field Effects : Neutral State

Calculate $\Delta G_j^{\text{rxn field}}(\text{PS}_i^o/\text{S}_i^o)$

$$\Delta G_j^{\text{rxn field}}(\text{PS}_i^o/\text{S}_i^o) =$$

$$\frac{1}{2}\left[\sum_{s(\text{PS}_i^o)}\sum_{jo}\frac{q_{jo}\sigma_s(\text{PS}_i^o)}{|r_{jo}-r_s|} - \sum_{s(\text{S}_i^o)}\sum_{jo}\frac{q_{jo}\sigma_s(\text{S}_i^o)}{|r_{jo}-r_s|}\right]$$



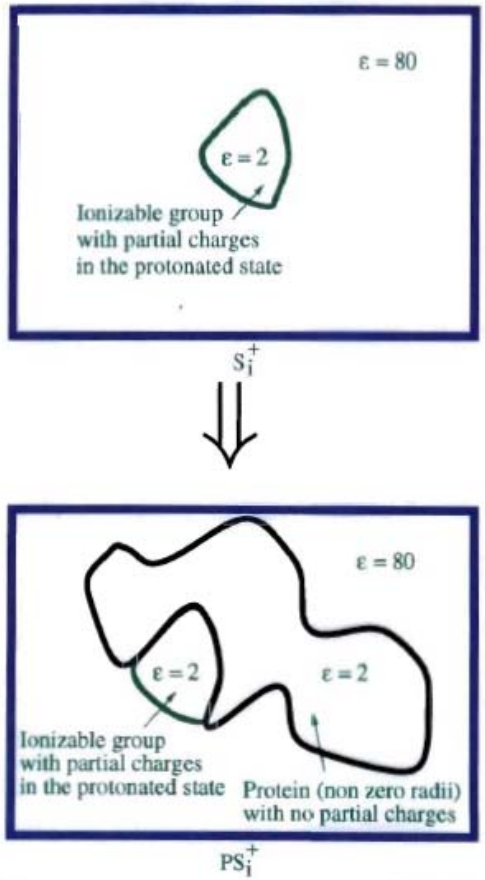
# Poisson-Boltzmann Calculations

## Reaction Field Effects : Ionized State

Calculate $\Delta G_j^{\text{rxn field}}(\text{PS}_i^+/\text{S}_i^+)$

$$\Delta G_j^{\text{rxn field}}(\text{PS}_i^+/\text{S}_i^+) =$$

$$\frac{1}{2}\left[\sum_{s(\text{PS}_i^+)}\sum_{j+}\frac{q_{j+}\sigma_s(\text{PS}_i^+)}{|r_{j+}-r_s|} - \sum_{s(\text{S}_i^+)}\sum_{j+}\frac{q_{j+}\sigma_s(\text{S}_i^+)}{|r_{j+}-r_s|}\right]$$

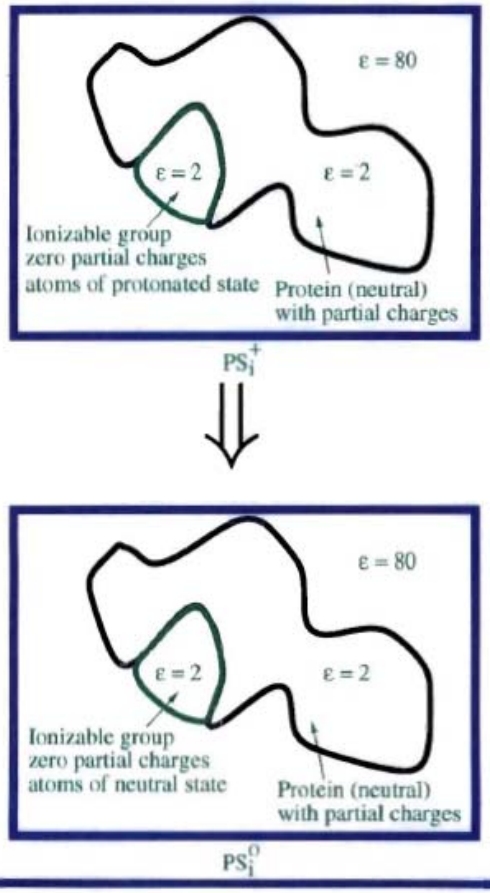


# Poisson-Boltzmann Calculations

## Permanent Dipole Effects

Permanent Dipole effects calculated based on sum of the effective potential at the atomic charge centers

$$\Delta\Delta G_j^{\text{perm dipole}} = \Delta G_j^{\text{perm dipole}}(\text{PS}_i^+/\text{S}_i^+) - \Delta G_j^{\text{perm dipole}}(\text{PS}_i^o/\text{S}_i^o)$$



# Overall Free Energy

$F =$

- **Potential**

(Scheraga & coworkers)

$$\sum_{ij \in NB} \epsilon_{ij}\left[\left(\frac{r^o_{ij}}{r_{ij}}\right)^{12} - \left(\frac{r^o_{ij}}{r_{ij}}\right)^{6}\right] +$$

$$\sum_{ij \in HB} \epsilon_{ij}\left[\left(\frac{r^o_{ij}}{r_{ij}}\right)^{12} - \left(\frac{r^o_{ij}}{r_{ij}}\right)^{10}\right] +$$

$$\sum_{ij \in ES} \frac{332\, q_i q_j}{D r_{ij}} + \sum_{k \in TOR} \frac{A_k}{2}(1 \pm \cos n_k \phi_k)$$

$E_{\text{vac}} +$

- **Entropic**

$$-\frac{k_B}{2} \ln\left[\text{Det}(\text{H}_{\text{vac},\gamma})\right]$$

$-TS_{\text{vac}} +$

- **Cavity**

(Honig & coworkers, 1988,1993,1995)


$\mathbf{F_{cavity} = \gamma(SA) + b}$

$F_{\text{cavity}} +$

- **Polarization**

(Honig & coworkers, 1988,1993,1995)


$\mathbf{F_{solv} = F_{polar}(\varepsilon=80) - F_{polar}(\varepsilon=1)}$

$F_{\text{polar}} +$

- **Ionization**

(Honig & coworkers, 1988,1993,1995)


$\mathbf{F_{ionize}(pH) = kT\ ln(Z)}$

$F_{\text{ionize}}$

# Probability of Helix Formation

Klepeis & Floudas 2002a

- Calculate probability of conformer *i* from free energy

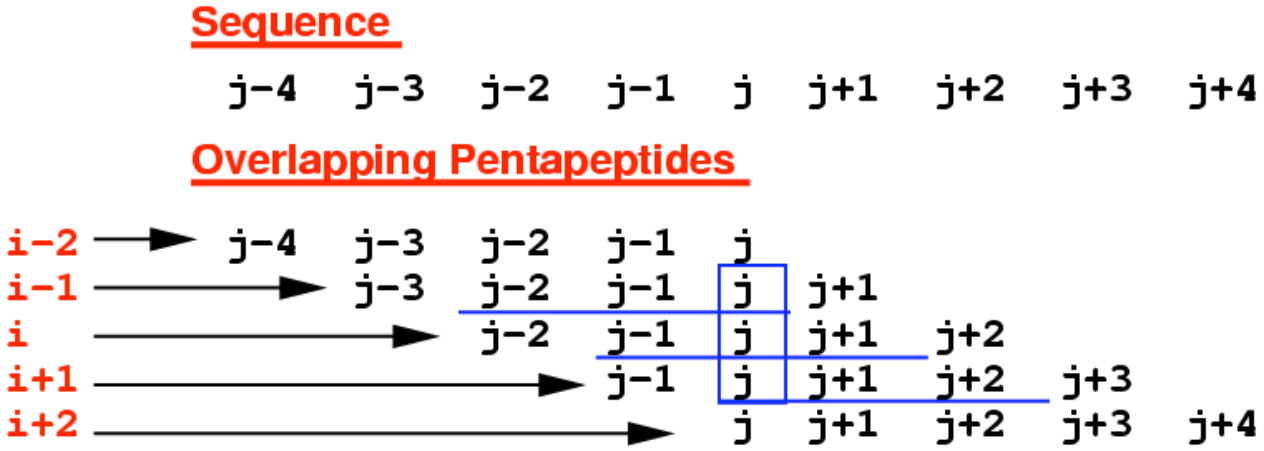
$$p_i = \frac{\exp[-\beta(F_o - F_i)]}{\sum_j \exp[-\beta(F_o - F_j)]}$$

- Cluster probabilities for helical (AAA) conformers

$$p_{\mathrm{AAA}} = \sum_{i \in \mathrm{AAA}} p_i$$

- Classify residues using probability of central peptides
- Probability calculation for residue *j* (pentapeptide) is

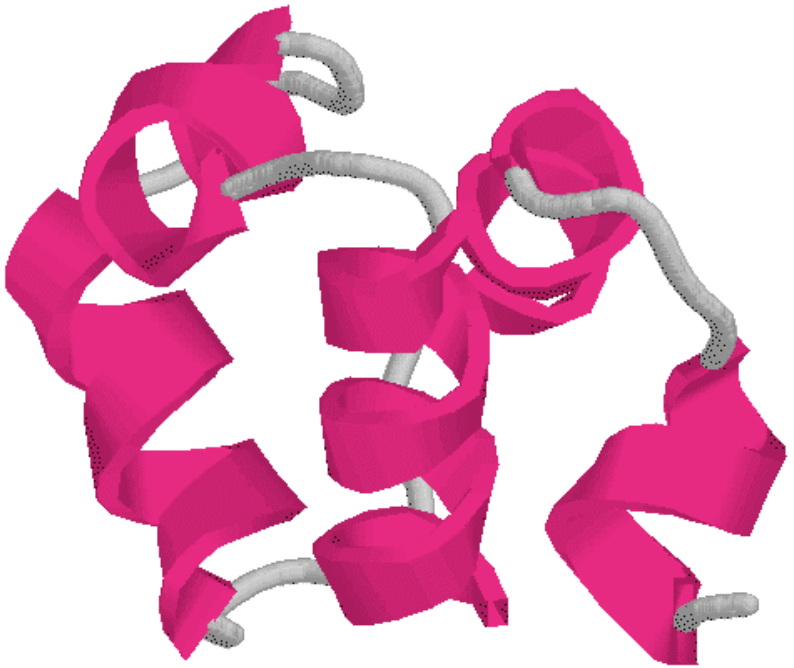
$$p^j_{\mathrm{AAA}} = \frac{p_{\mathrm{AAA},i-1} + p_{\mathrm{AAA},i} + p_{\mathrm{AAA},i+1}}{3}$$



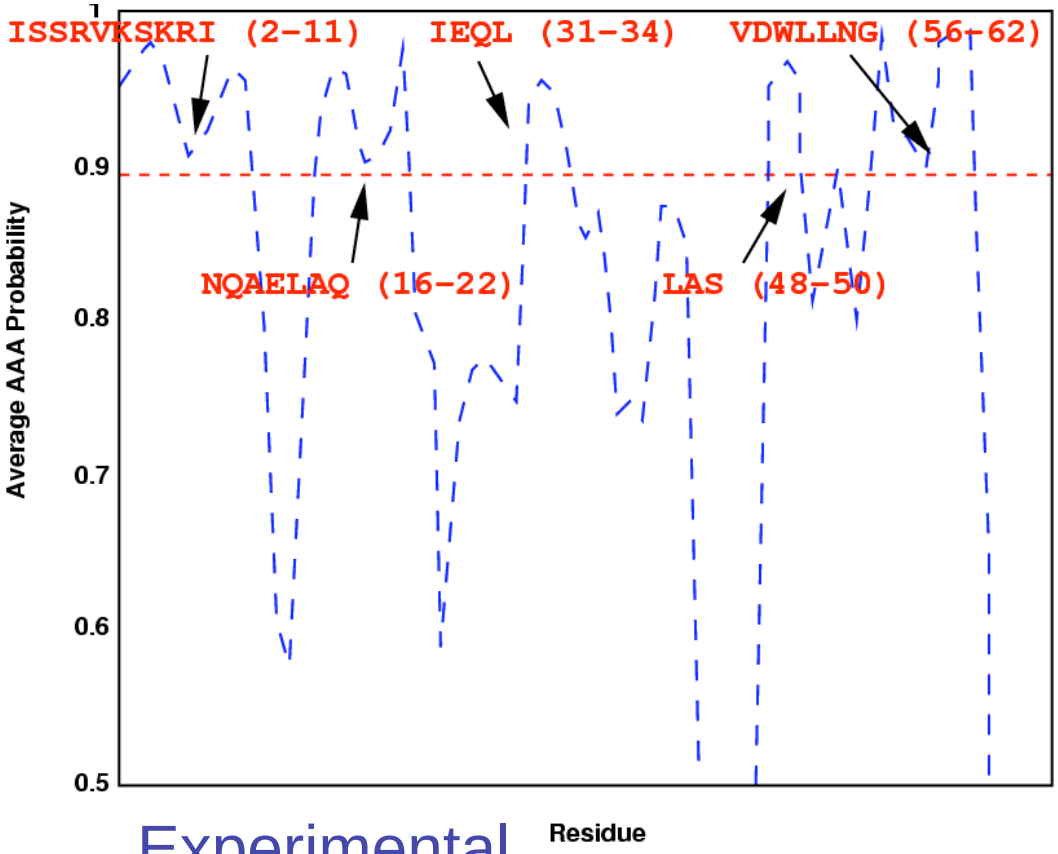
# Computational Study : 1R69

## N-terminal domain of phage 434 repressor protein

- 69 residue fragment of N-terminal domain
- Dimer involved with operator sites in phage genome
- N-terminal domain binds to DNA
- Compact hydrophobic interior
- Five helices
- Extended first helix
- Helix 2 and 3 form helix-turn-helix motif



# Computational Study : 1R69




| Predicted | Exp |
|---|---|
| 2-11 | 1-12 |
| 16-22 | 16-22 |
| 31-34 | 28-35 |
| 48-50 | 45-50 |
| 56-62 | 56-64 |

## Experimental

SISSRVKSKRIQLGLNQAELAQKVGTTQQSIEQLENGKTKRPRFLPELASALGVSVDWLLNGTSDSNVR

- Helices from 1-12, 16-22, 28-35, 45-50, 56-64

## Comparison with PSIPRED

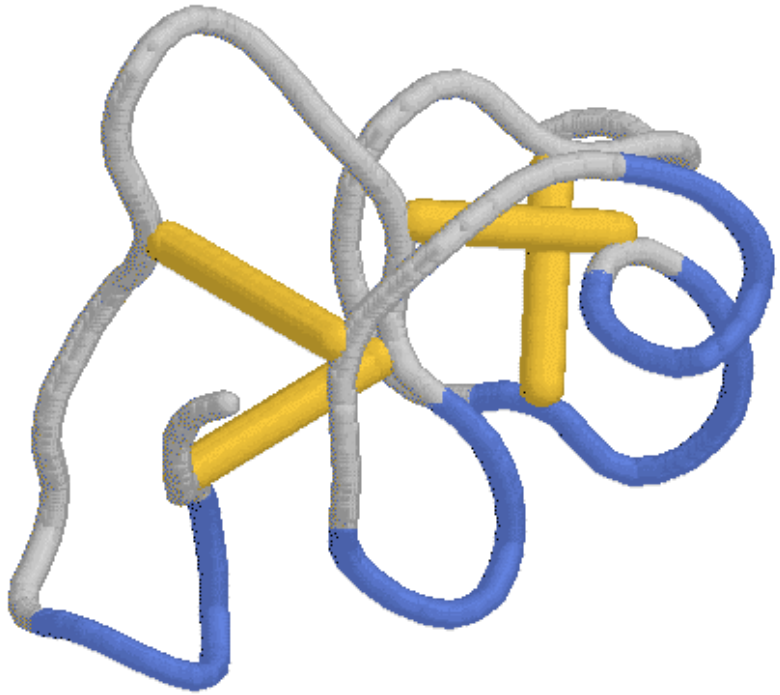
SISSRVKSKRIQLGLNQAELAQKVGTTQQSIEQLENGKTKRPRFLPELASALGVSVDWLLNGTSDSNVR

- Helices from 2-12, 17-24, 28-35, 42-52, 56-59

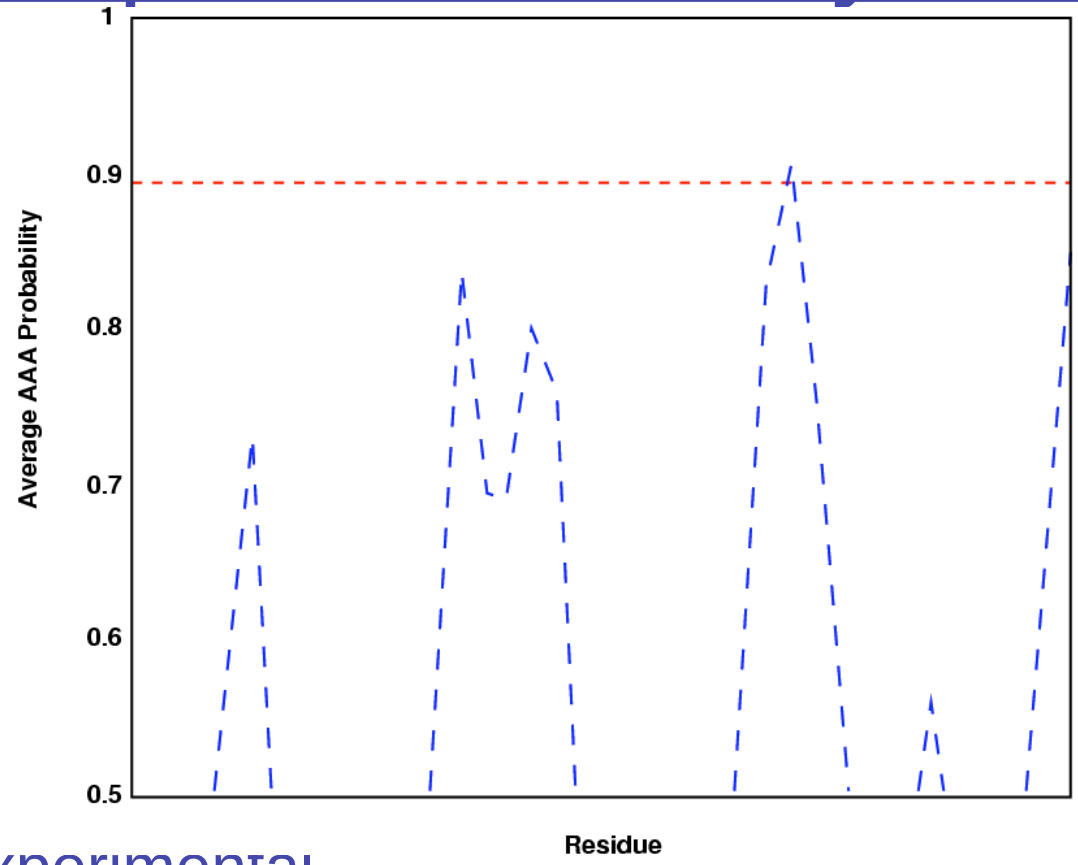
# Computational Study : 9WGA

## Wheat germ agglutinin

- Dimeric plant lectin
- 171 residue chains with four domains each (A, B, C and D)
- Internal repeating domain of 43 residues
- Homology in all Cys and many Gly positions
- Four disulfide bridges per domain
- Tight fold
- No helical or beta structure



# Computational Study : 9WGA




Experimental

KRCGSQAGGATCPNNHCCSQYGHCGFGAEYCGAGCQGGPCRAD

- Helices from 1-12, 16-22, 28-35, 45-50, 56-64

Comparison with PSIPRED

KRCGSQAGGATCPNNHCCSQYGHCGFGAEYCGAGCQGGPCRAD

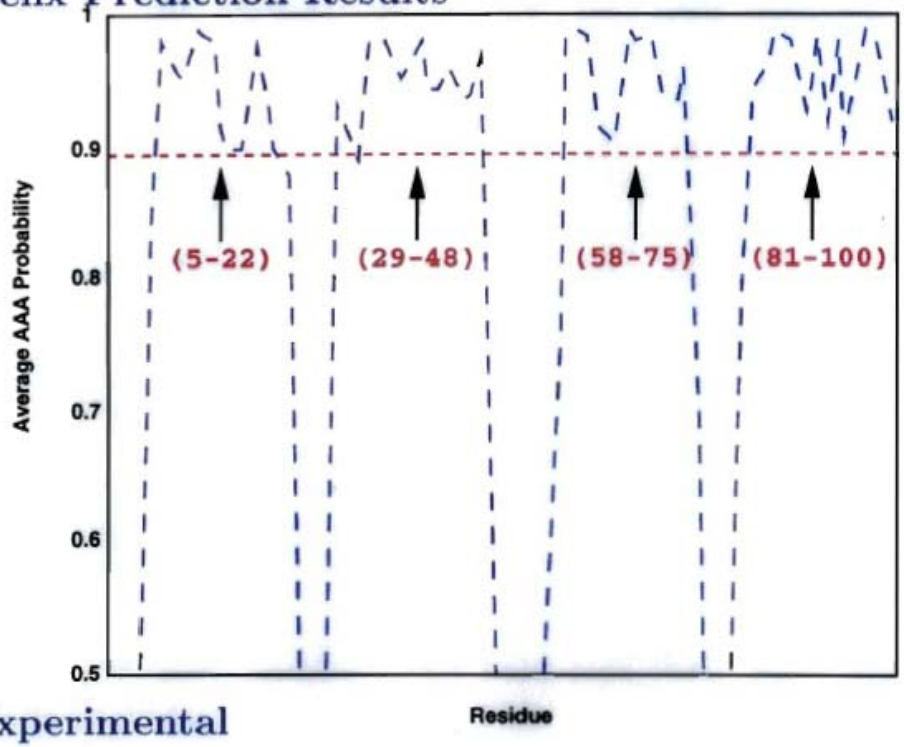
- Helices from 2-12, 17-24, 28-35, 42-52, 56-59

# Computational Study : Blind Test

- 102 residue sequence
  (Professor Michael Hecht, Princeton University)
- No knowledge of secondary/tertiary structure
- Experimental analysis in progress
- Designed as four-helix bundle protein

# Computational Study : Blind Test

## Helix Prediction Results




## Experimental

- Four-helix bundle expected

Agreement with design goals

## Comparison with PSIPRED

- Helices from 5-21, 30-49, 56-75, 80-100

# Structure Prediction In Protein Folding: Outline

- Introduction to Protein Structure Prediction
- Free Energy Calculations in Oligo-peptides
- Prediction of Helical Segments
- Prediction of Beta Sheet Topologies
- Prediction of Loop Structures
- Derivation of Restraints
- Prediction of Protein Tertiary Structure

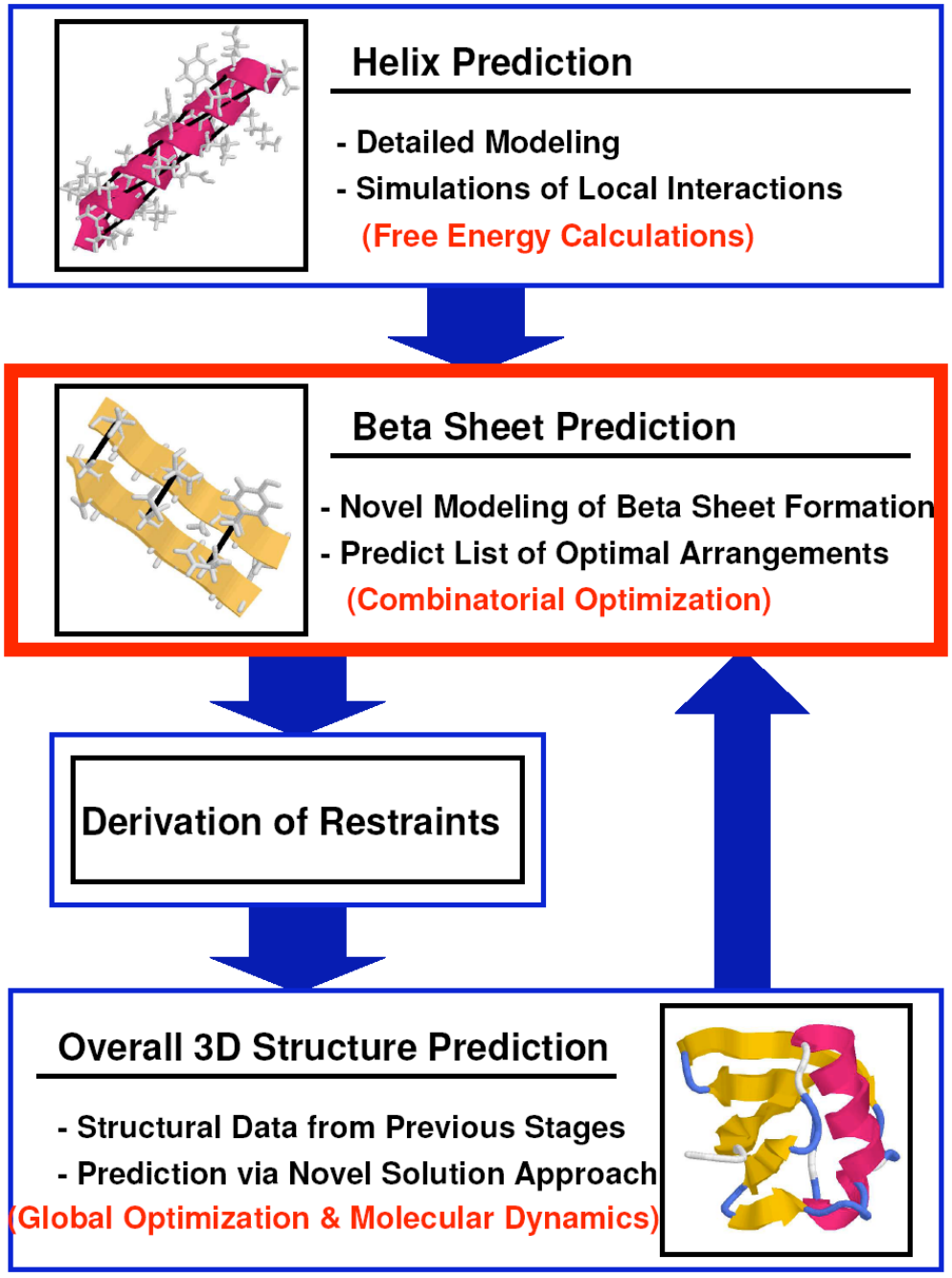
# Prediction of Beta Sheet Topologies via Integer Linear Optimization

## Relevant References:


- **Klepeis J.L. and C.A. Floudas, "Prediction of Beta-Sheet Topology and Disulfide Bridges in Polypeptides", Journal of Computational Chemistry, 24, 191-208 (2003).**

# ASTRO-FOLD

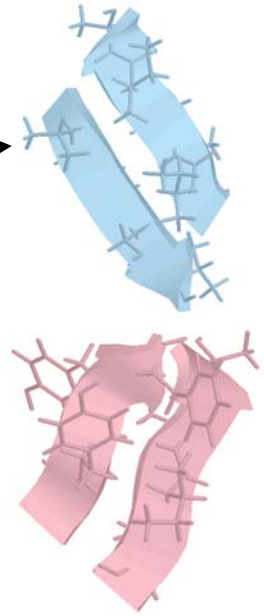
Klepeis & Floudas, 2002c



# Formation of β-Sheets

## Major challenge for accurate structure prediction

- Prediction of β-strand location not accurate
- No reliable method for β-sheet topology
  - Antiparallel β-sheets
  - Parallel β-sheets
- How to treat formation of disulfide bridges



## Physical Understanding

- Local forces not as dominant (as for helices)
- Non local forces are significant
  - Hydrophobic collapse
  - Tertiary contacts

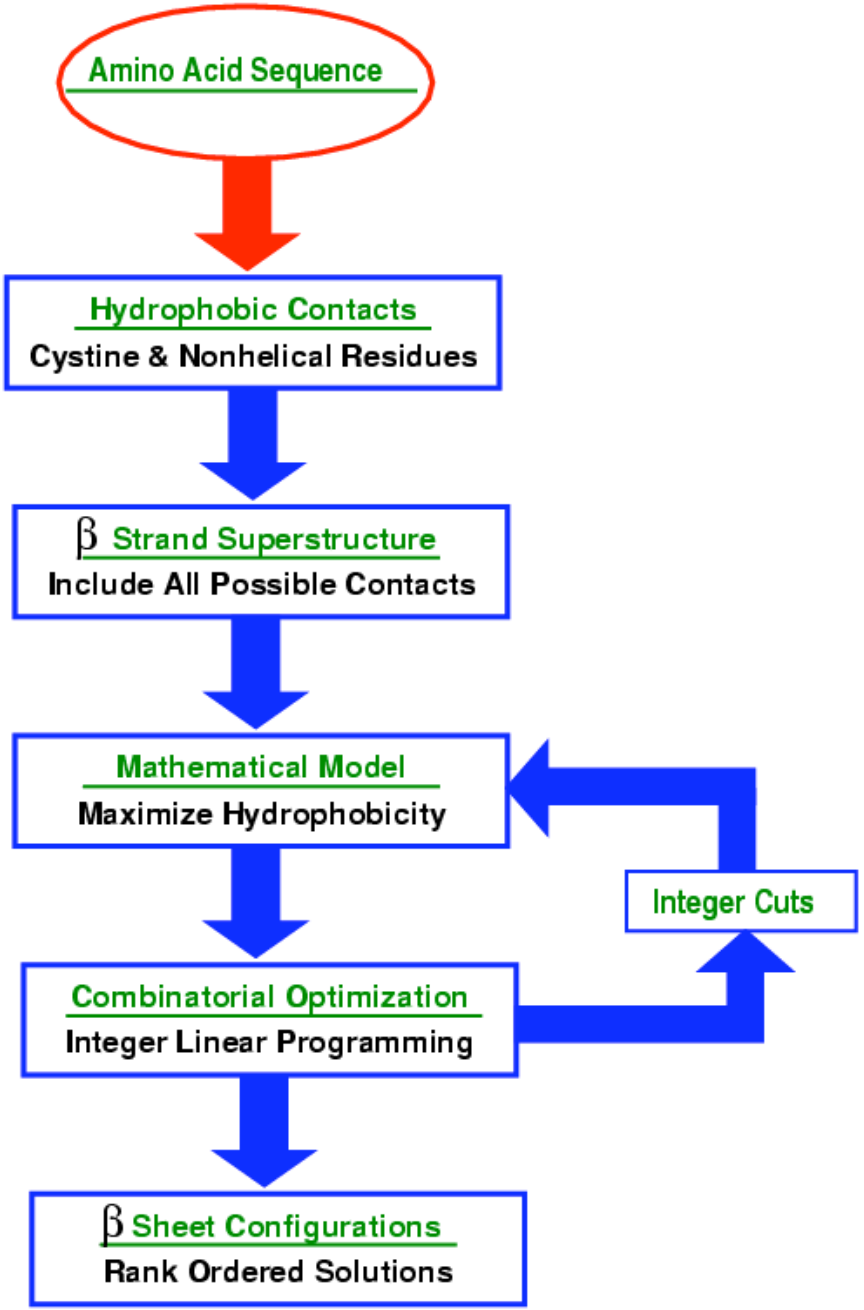
## Experimental Evidence

- Hydrophobic collapse proceeds rapidly

**Hydrophobic forces drive β-sheet formation**

# β-Sheet Prediction

Klepeis & Floudas, 2002b



# β-Strand Protocol

- Classify nonhelical and all cystine residues

| | | |
|---|---|---|
| Hydrophobic | H : | *Leu,Ile,Val,Phe,Met,Cys,Tyr,Trp* |
| Bridge | B : | *Ala,Thr* |
| Turn | T : | *Asn,Asp,Gly,Pro,Ser* |
| Other | N : | *Arg,Lys,Glu,Gln,His* |

- Scan sequence for Hydrophobic residues and identify Hydrophobic to Hydrophobic segments
- Build β strands using rules for intervening residues

NNBTHHBNHBTHTBHHBHTNTN

- Scan sequence and identify Turn to Turn segments
- Modify β strands which enclose, intersect or are enclosed within the Turn to Turn segments

NNBTHHBNHBTHTBHHBHTNTN

93 % strand prediction accuracy for 11000 strands from over 2000 PDB sequences (50 to 150 amino acids)

# Full $\beta$ Strand Protocol

## Residue Classification

Classify all residues not belonging to $\alpha$-helices as *hydrophobic*, *bridge*, *turn* or *other* residues.

- The set of residues, $\mathcal{H}$, are considered to be *hydrophobic* :

$$\mathcal{H} = \{\text{Leu,Ile,Val,Phe,Met,Cys,Tyr,Trp}\}$$

- The set of residues, $\mathcal{B}$, are considered to be *bridge* :

$$\mathcal{B} = \{\text{Ala,Thr}\}$$

- The set of residues, $\mathcal{T}$, are considered to be *turn* :

$$\mathcal{T} = \{\text{Asn,Asp,Gly,Pro,Ser}\}$$

- The set of residues, $\mathcal{N}$, are considered to be *other* :

$$\mathcal{N} = \{\text{Arg,Lys,Glu,Gln,His}\}$$

For the case of a Ser residue juxtaposed to a residue belonging to the set of bridge residues, $\mathcal{B}$, the classification of the individual Ser residue is changed from $\mathcal{T}$ to $\mathcal{B}$.

# Illustration of β–strand Superstructure

Bovine Pancreatic Trypsin Inhibitor

**Position of experimentally determined strands : O**

**Position of PSIPRED predicted strands : E**

```
1234567890  1234567890  1234567890  1234567890  1234567890
RPDFCLEPPY  TGPCKARIIR  YFYNAKAGLC  QTFVYGGCRA  KRNNFKSAED
CCCCCCCCCC  CCCCCCCEEE  EEEECCCCEE  EEEEECCCCC  CCCCCCCHHH
----------  -------OOO  OOOO----OO  OOOOO-----  ----------

12345678
CMRTCGGA
HHHHHCCC
--------
```

# Illustration of β–strand Superstructure

Bovine Pancreatic Trypsin Inhibitor

**Position of predicted strands : X**

**Position of cystines : S**

**Position of experimentally determined strands indicated in red**

| 1234567890 | 1234567890 | 1234567890 | 1234567890 | 1234567890 |
|---|---|---|---|---|
| RPDFCLEPPY | TGPCKARIIR | YFYNAKAGLC | QTFVYGGCRA | KRNNFKSAED |
| N----HNTTH | BTTHNBNHHN | HHHTBNBTHH | NBHHHTTHNB | NNTTHN---- |
| ----S----- | ---S--XXXX | XXX-----XS | XXXXX--S-- | ---------- |

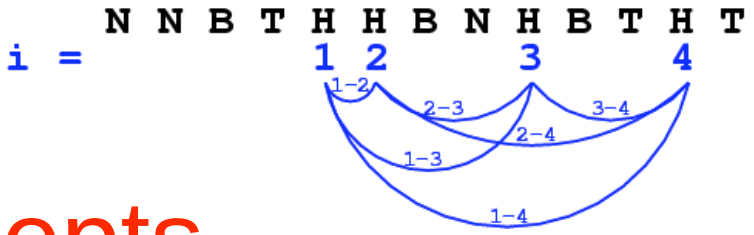
| 12345678 |
|---|
| CMRTCGGA |
| ----HTTB |
| S---S--- |

# Residue-based Concepts

- Identify set *i* and assign hydrophobicity index $H_i$ to nonhelical Hydrophobic and all cysteine residues

```
          N N B T H H B N H B T H T B H H B
i =             1 2     3     4       5 6
```

- Assign binary variable $y_{ij}$ to each possible unique Residue-to-Residue contact

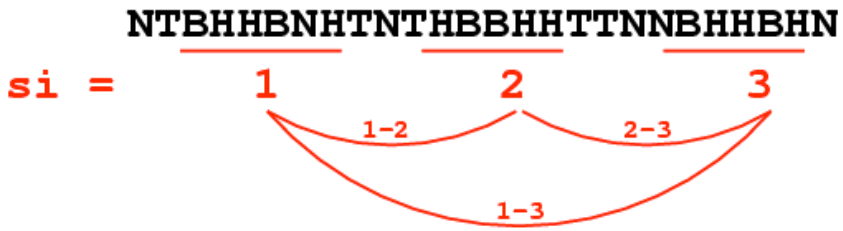



# Strand-based Concepts

- Identify set *si* and assign hydrophobicity weight $S_{si}$ according to superstructure of potential β-strands

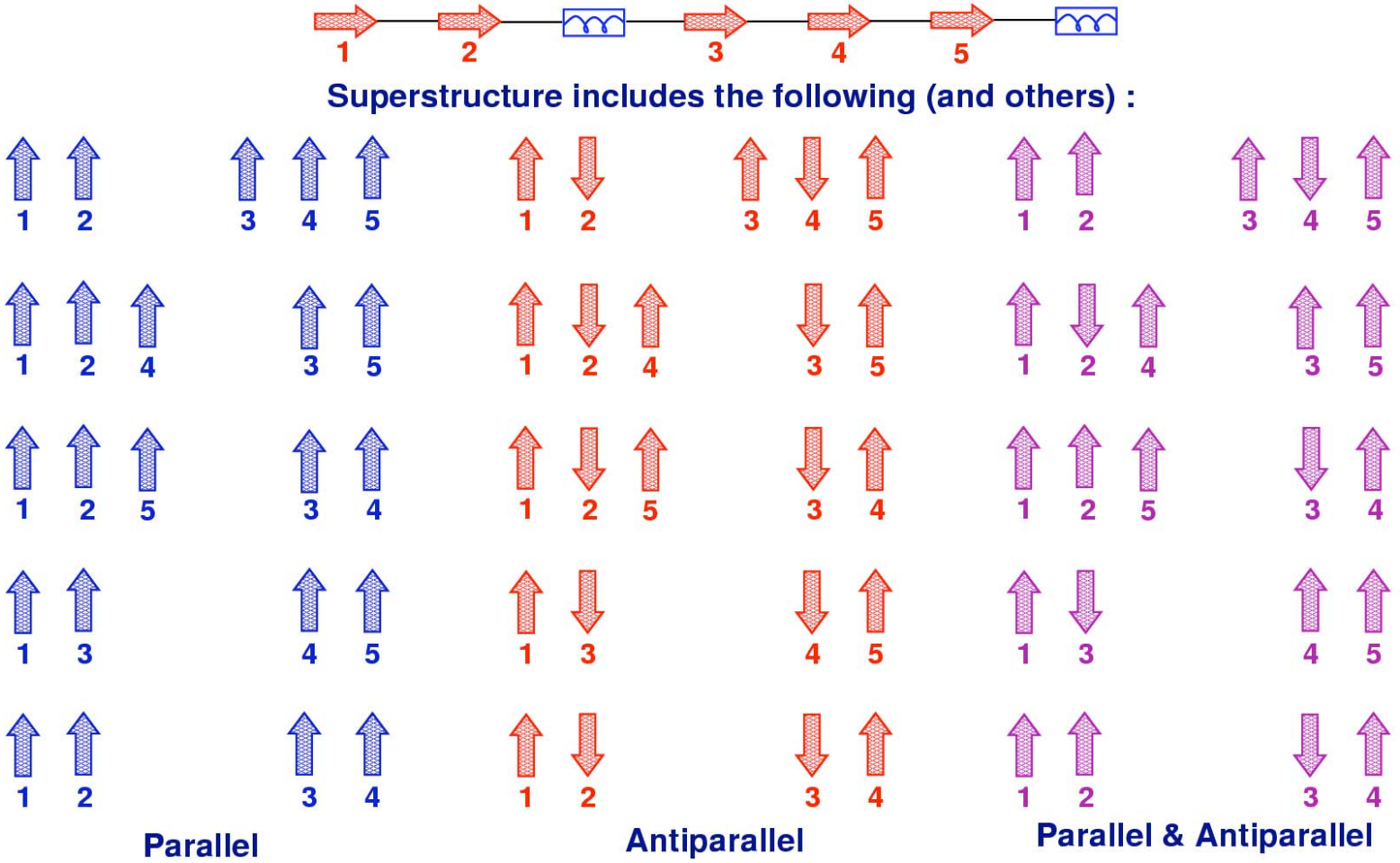
```
          NTBHHBNHTNTHBBHHTTNNBHHBHN
si =      1       2       3
```

- Assign binary variable $w_{si,sj}$ to each possible unique Strand-to-Strand contact



# β-strand Superstructure

- Model incorporates ALL possible alternatives




- Number of postulated strands may be greater than actual
- Many possible β–sheet arrangements are allowable

# Formulation : Key Concepts

Klepeis & Floudas 2002b

## Binary variables

0-1 variables are used to characterize residue-to-residue and strand-to-strand contacts

## Linear objective function

Objective is to maximize the hydrophobic potential as controlled by the binary variables

## Linear constraints

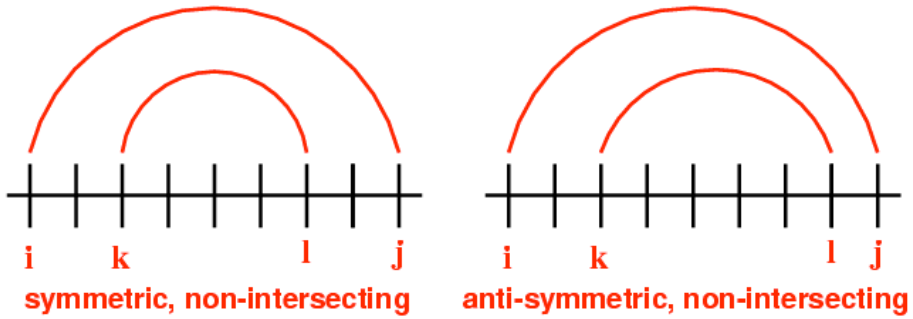
Constraints account for different combinations of residue and strand contacts (e.g., parallel/antiparallel)

## Integer cuts

Iterative addition of these constraints allow for the generation of a ranked list of optimal solutions

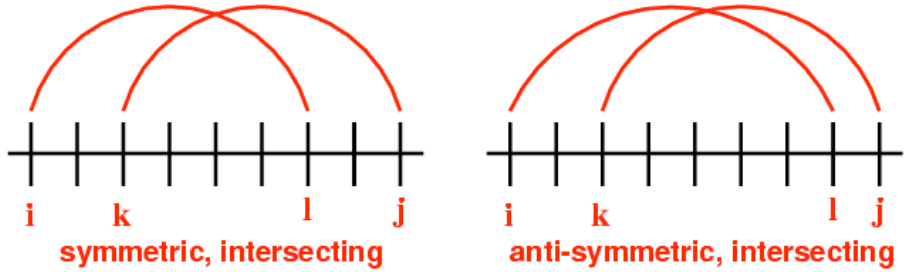
# Constraint Functions

- Allowable antiparallel combinations



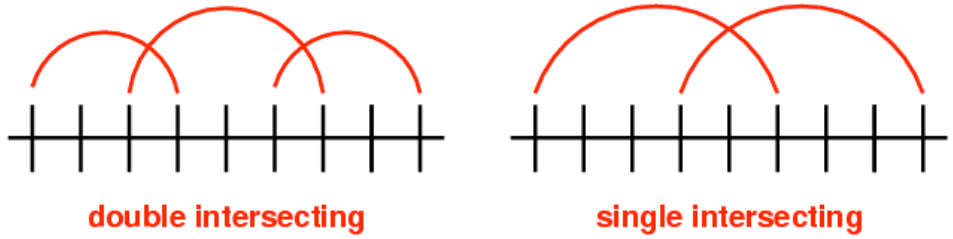

$$y_{ij} + y_{kl} \leq 1 \quad \forall \quad P(i) + P(j) \neq P(k) + P(l)$$

- Allowable parallel combinations




$$y_{ij} + y_{kl} \leq 1 \quad \forall \quad P(k) - P(i) \neq P(l) - P(j)$$

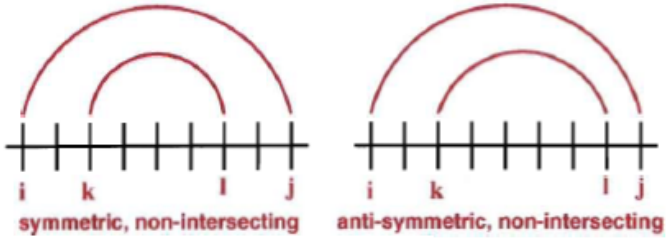
- Limit number of strand contacts to (2)
- Disallow extended β-ladders
- Disallow double intersecting loops



# Constraint Functions

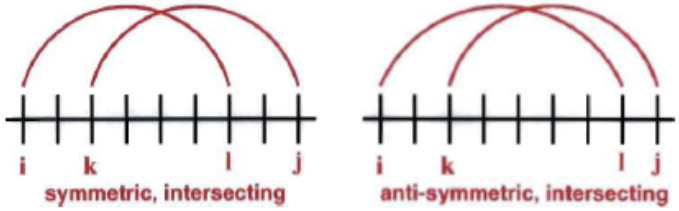
- Allowable antiparallel combinations

$$y_{ij} + y_{kl} \leq 1 \quad \forall \quad P(i) + P(j) \neq P(k) + P(l)$$

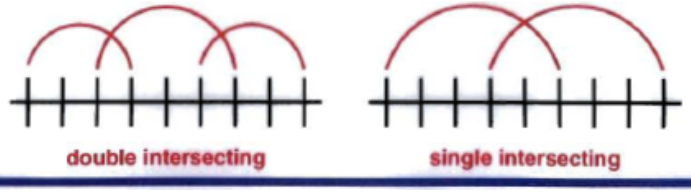



- Allowable parallel combinations

$$y_{ij} + y_{kl} \leq 1 \quad \forall \quad P(k) - P(i) \neq P(l) - P(j)$$



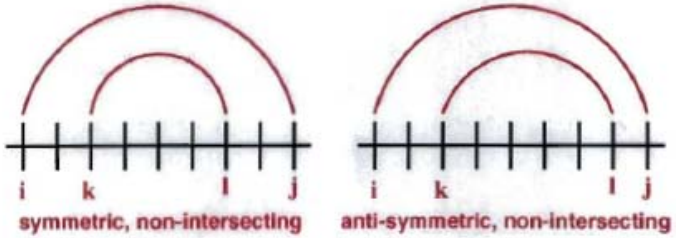

- Limit number of strand contacts to $NS_{si} = 2$
- Disallow extended $\beta$ ladders
- Disallow double intersecting loops



# Constraint Functions (Residue)

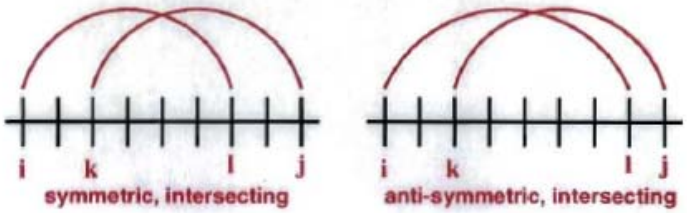
- Allowable antiparallel combinations $(y_{ij})$

$$
\begin{array}{rcl}
y_{ij} + y_{kl} & \leq & 1 \\
\forall \quad P(i) + P(j) & \neq & P(k) + P(l) \\
y_{ij} \quad \text{OR} \quad y_{kl} & \notin & \{\text{Cys,Cys}\}
\end{array}
$$




- Allowable parallel combinations $(y_{ij})$

$$
\begin{array}{rcl}
y_{ij} + y_{kl} & \leq & 1 \\
\forall \quad P(k) - P(i) & \neq & P(l) - P(j) \\
y_{ij} \quad \text{OR} \quad y_{kl} & \notin & \{\text{Cys,Cys}\}
\end{array}
$$



# Constraint Functions (Strand)

- Limit number of strand contacts to $NS_{si} = 2$ $(w_{si,sj})$

$$\sum_{sj,Q(si)<Q(sj)} w_{si,sj} + \sum_{sj,Q(sj)<Q(si)} w_{sj,si} \leq NS_{si} \quad \forall \; si$$

- Disallow extended $\beta$ ladders $(w_{si,sj})$

$$\sum_{sj,Q(si)\leq Q(sj)\leq Q(si)+2} \sum_{sk,Q(sk)=Q(sj)+1} w_{sj,sk} \leq 2 \quad \forall \; si$$

- Disallow more than one strand-to-strand match from two consecutive strands on one side of strand $si$ $(w_{si,sj})$

$$\sum_{sj,Q(sj)-3<Q(si)<Q(sj)} w_{si,sj} \leq 1 \quad \forall \; si$$
$$\sum_{sj,Q(si)-3<Q(sj)<Q(si)} w_{sj,si} \leq 1 \quad \forall \; si$$

# Constraint Functions (Strand)

- Disallow strand-to-strand contact - generic $(w_{si,sj})$

$$w_{si,sj} \leq DS_{si,sj}$$

- Disallow double intersecting loops $(w_{si,sj})$

$$w_{sm,sn} \leq 2 - w_{si,sj} - w_{sk,sl}$$

$\forall \quad Q(sj) > Q(si)\,,\ Q(sl) > Q(sk)\,,\ Q(sn) > Q(sm)\,,$
$Q(sk) > Q(si)\,,\ Q(sl) > Q(sj)$ AND
Condition 1 OR 2 OR 3 OR 4 OR 5 OR 6 OR 7 is TRUE

Condition 1 : $Q(sm) > Q(si)\,,\ Q(sm) < \text{MIN}[Q(sj), Q(sk)]\,,$
$Q(sn) > \text{MAX}[Q(sj), Q(sk)]\,,\ Q(sn) < Q(sl)$

Condition 2 : $Q(sm) > Q(si)\,,\ Q(sm) < \text{MIN}[Q(sj), Q(sk)]\,,$
$Q(sn) > \text{MAX}[Q(sj), Q(sk)]\,,\ Q(sn) = Q(sl)$

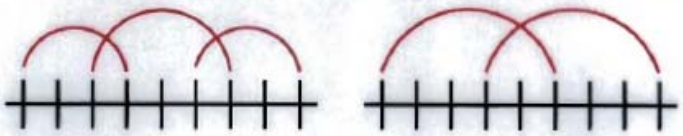
Condition 3 : $Q(sm) = Q(si)\,,\ Q(sm) < \text{MIN}[Q(sj), Q(sk)]\,,$
$Q(sn) > \text{MAX}[Q(sj), Q(sk)]\,,\ Q(sn) = Q(sl)$

Condition 4 : $Q(sm) = Q(si)\,,\ Q(sn) = Q(sl)$

Condition 5 : $Q(sm) = Q(sk)\,,\ Q(sn) = Q(sj)$

Condition 6 : $Q(sm) = Q(sk)\,,\ Q(sn) \leq Q(sj)$

Condition 7 : $Q(sm) \geq Q(sk)\,,\ Q(sn) \leq Q(sj)$



# Objective Function

$$\max \quad \sum_i \sum_{j,P(i)+2<P(j)} (H_i + H_j + H_{ij}^{\text{add}}) y_{ij}$$

$$+ \quad \sum_{si} \sum_{sj,Q(si)<Q(sj)} (S_{si} + S_{sj}) w_{si,sj}$$

$$y_{ij} = \begin{cases} 1 & \text{if} \quad i,j \quad \text{form contact} \\ 0 & \text{if} \quad i,j \quad \text{do not form contact} \end{cases} \qquad \forall i < j$$

$$w_{si.sj} = \begin{cases} 1 & \text{if} \quad si,sj \quad \text{form contact} \\ 0 & \text{if} \quad si,sj \quad \text{do not form contact} \end{cases} \qquad \forall si < sj$$

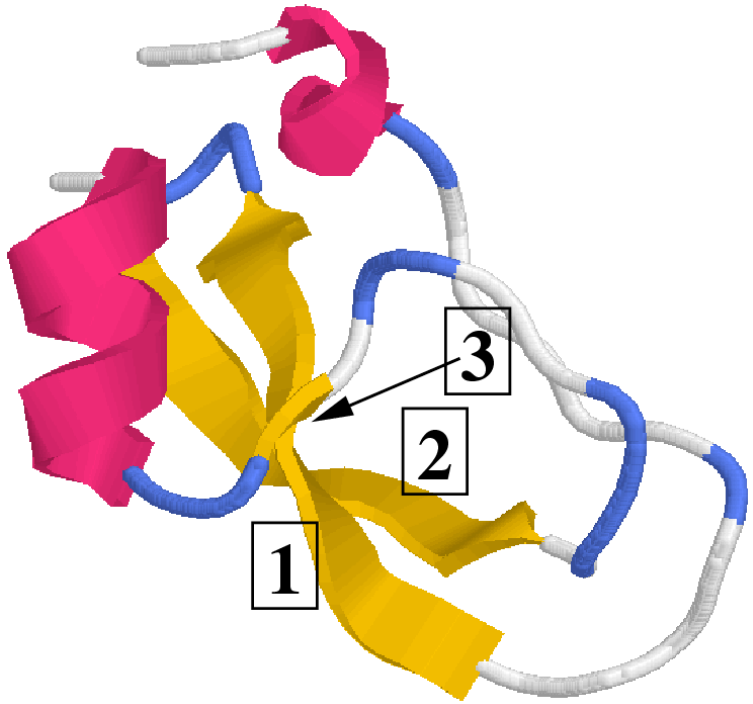
- Maximization of hydrophobic potential
- Additional disulfide contact energy

$$H_{ij}^{\text{add}} = \begin{cases} \frac{\sum_{k,P(i)\le P(k)\le P(j)} H_k}{P(j)-P(i)} & \text{if} \quad \{i,j\} \in \{\text{Cys}\} \\ 0 & \text{otherwise} \end{cases}$$

# Computational Study of β Prediction

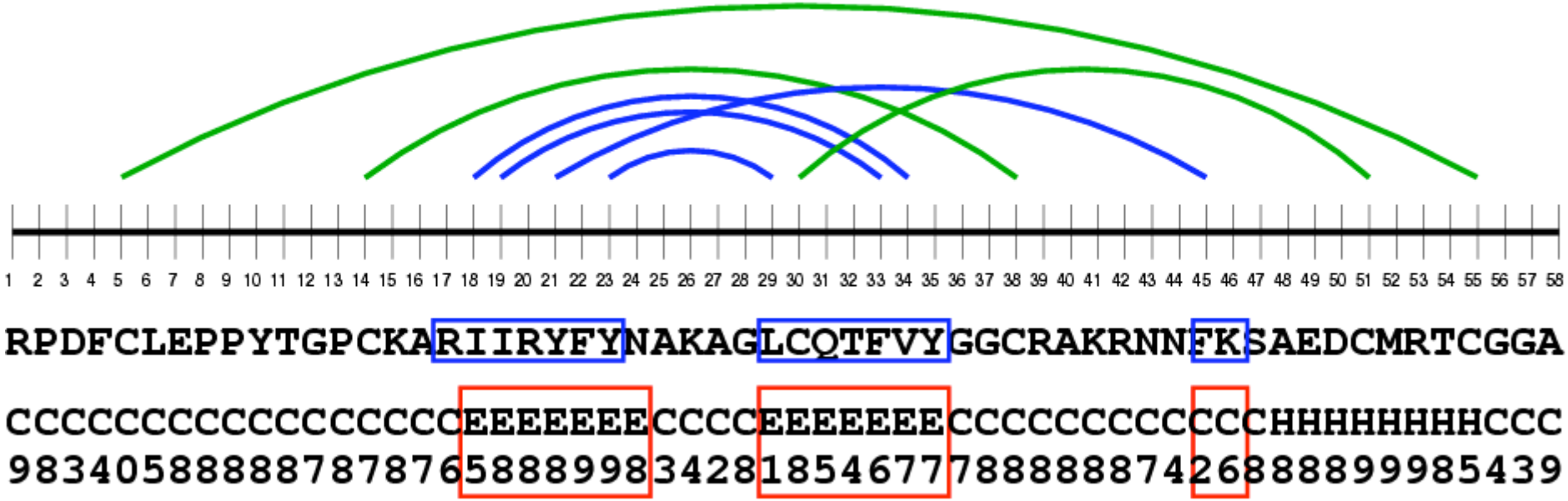
## Bovine Pancreatic Trypsin Inhibitor

- 58 residue protein
- Inhibits serine proteases
- Three disulfide bonds
  - Cys5-Cys55, Cys14-Cys38, Cys30-Cys51
- Two antiparallel strands (Strand 1 to Strand 2)
- Hydrophobic residue match (Strand 1 to Strand 3)



# Computational Study of β Prediction

## Bovine Pancreatic Trypsin Inhibitor




- Disulfide bridge matches : 5-55, 14-38, 30-51
- Residue matches (Strand 1 to Strand 2) :
  18-34 (Ile-Val), 19-33 (Ile-Phe), 23-29 (Tyr-Leu)
- Residue match (Strand 1 to Strand 3) : 22-45 (Phe-Phe)

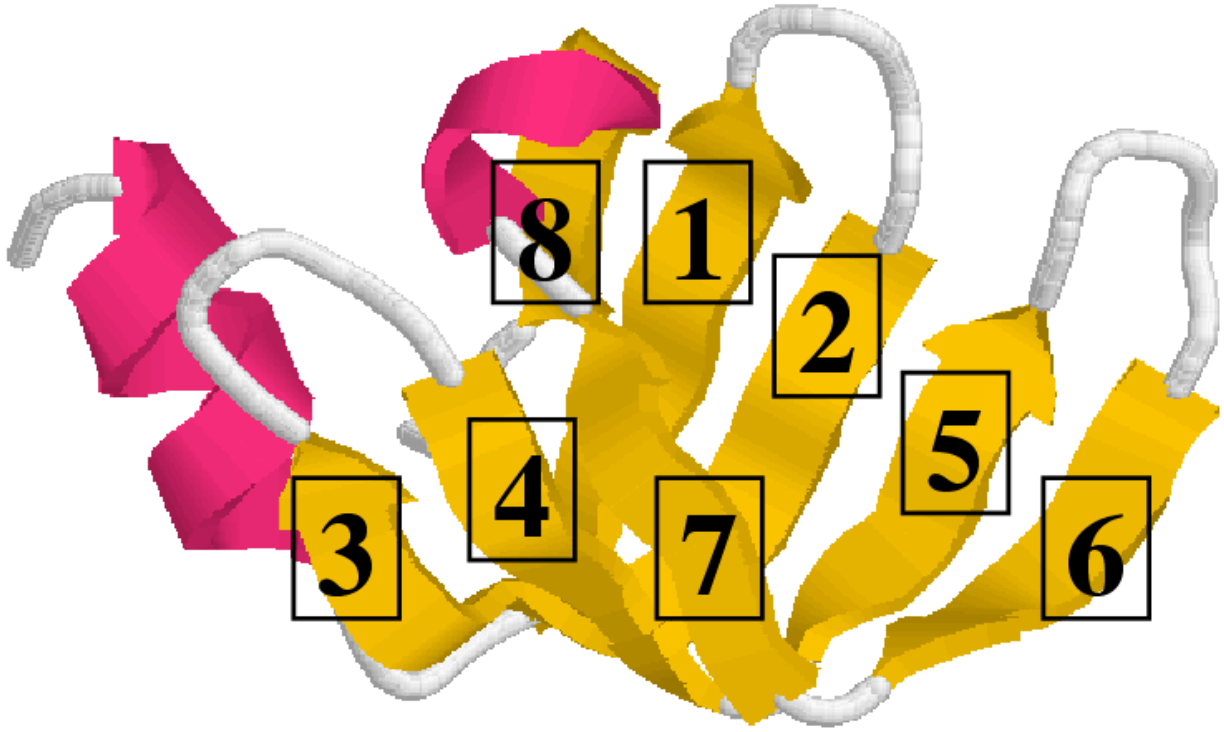
PSIPRED comparison

- 2 strands :19-24 (79 %)  and 29-35 (54 %)
- β-sheet configuration can NOT be identified

# Computational Study of β Prediction

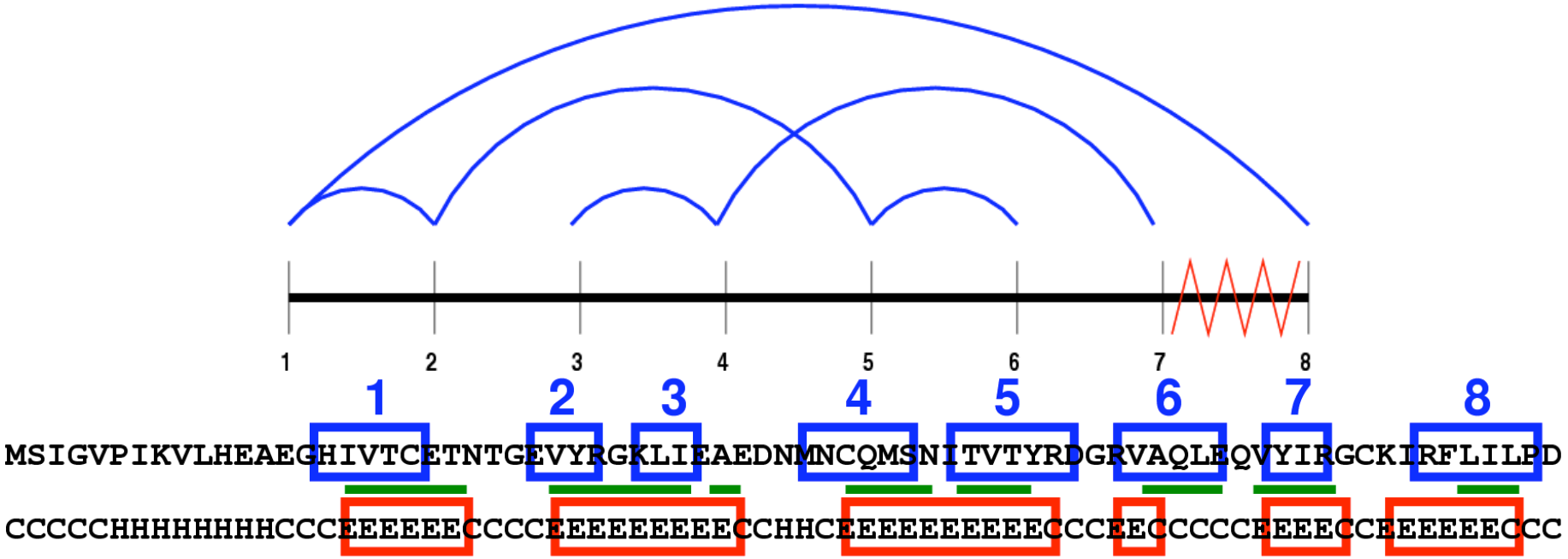
## SmD3 : Small Nuclear Ribonucleoprotein (T0059)

- 75 residue protein
- Common fold of SH3 proteins
- N-terminal helix
- Set of antiparallel β-sheets
- Barrel-like topolgy



# Computational Study of β Prediction

SmD3 : Small Nuclear Ribonucleoprotein (T0059)

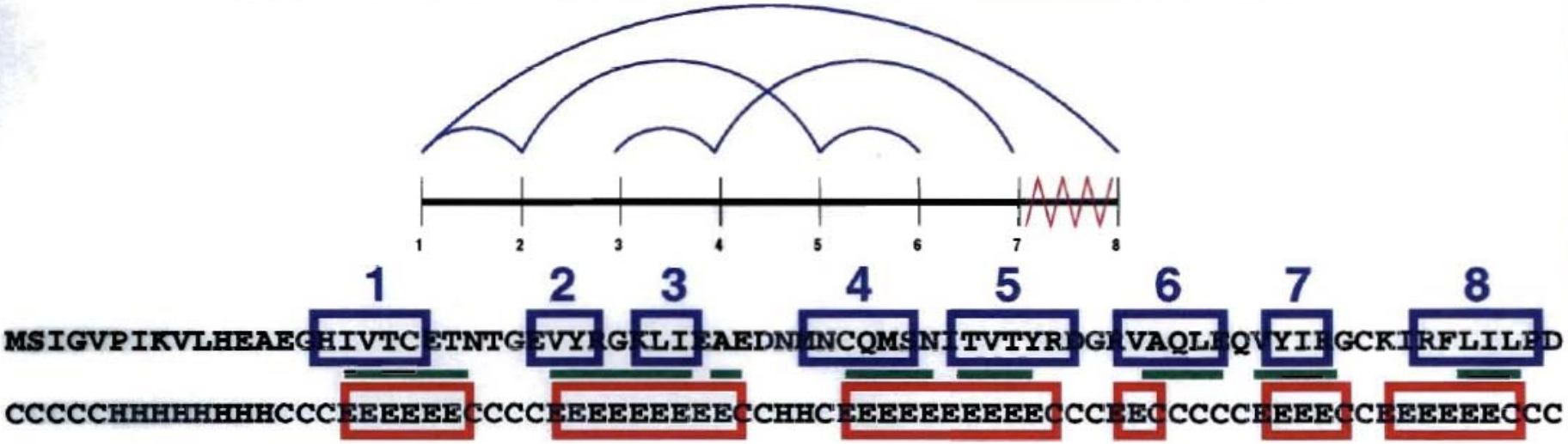



- Multiple global optima with six contacts
- Consistent features include matches between strands 1-2 and 1-8
- One of the six global optima corresponds to experimental observations

PSIPRED comparison

- Length of β strands inconsistent with experimental results
- β-sheet configuration can NOT be identified

# Computational Study : T0059




- Multiple global optima with six contacts
- One global optimum provides exact agreement with experiment

| Optimum | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| Match 1 | 1-2 | 1-2 | 1-2 | 1-2 | 1-2 | 1-2 | 1-2 |
| Match 2 | 1-8 | 1-8 | 1-8 | 1-8 | 1-8 | 1-8 | 1-8 |
| Match 3 | 2-5 | 2-6 | 2-4 | 2-4 | 2-4 | 2-4 | 2-5 |
| Match 4 | 3-4 | 3-4 | 3-7 | 3-6 | 3-5 | 3-5 | 3-4 |
| Match 5 | 5-6 | 4-5 | 4-5 | 4-5 | 4-6 | 4-7 | 4-6 |
| Match 6 | 4-7 | 5-7 | 5-6 | 5-7 | 5-7 | 5-6 | 5-7 |

## PSIPRED Comparison

- Length of $\beta$ sheets inconsistent with experimental results
- $\beta$ sheet configuration can NOT be identified

# Structure Prediction In Protein Folding: Outline

- Introduction to Protein Structure Prediction
- Free Energy Calculations in Oligo-peptides
- Prediction of Helical Segments
- Prediction of Beta Sheet Topologies
- Prediction of Loop Structures
- Derivation of Restraints
- Prediction of Protein Tertiary Structure

# Prediction of Loop Structures with Flexible Stems

**Relevant References:**


- **Klepeis J.L. and C.A. Floudas, "Analysis and Prediction of Loop Segments in Protein Structures", Computers and Chemical Engineering, 29, 423-436 (2005).**
- **Monnigmann M. and C.A. Floudas, "Protein Loop Structure Prediction with Flexible Stem Geometries", Proteins, 61, 748-762 (2005).**

# Outline

## Protein Loop Prediction

- Motivational Examples
- Problem Definition
  - Current Approaches
  - Limitations
- Loop Prediction Strategies
  - Overlapping Oligopeptides
  - Pivot Point Constraints
- Computational Studies
  - Bovine Pancreatic Trypsin Inhibitor
  - Immunoglobulin Binding Domain of Protein G
  - T0114 : Antifungal Protein (*Streptomyces Tendae*)
- Conclusions

# Functional Significance of Loops

## Endotoxins

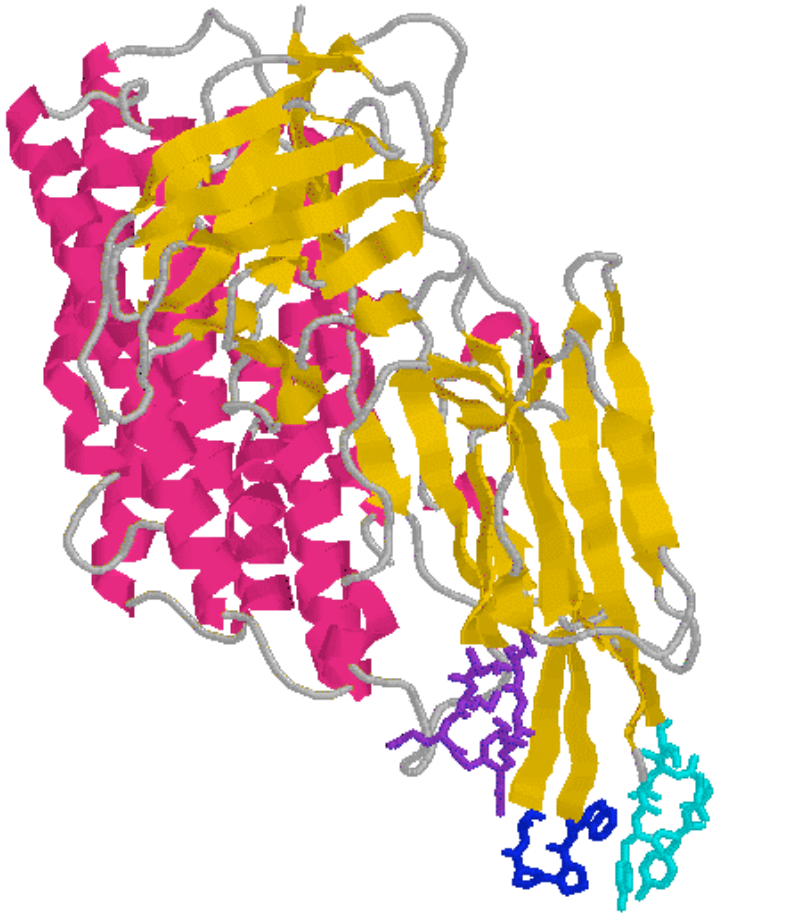
- Natural pesticide used in agriculture
- CryIIIA class is active against potato beetle

## Function

- Bind to receptor proteins
- Inset into membrane
- Function as ion channels

## Structure-Function

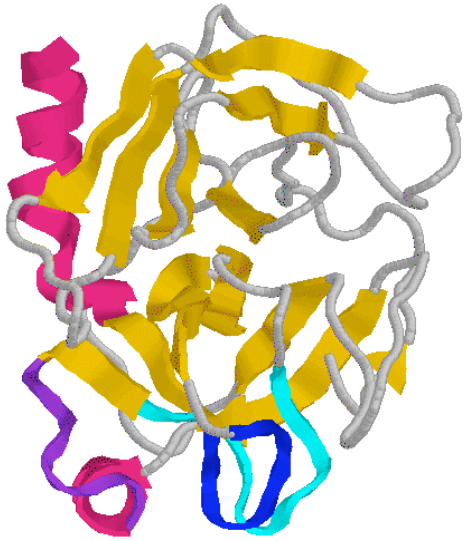
- 3 surface loops involved in receptor binding & specificity
- Alanine replacements affect receptor binding (loops 1 & 3), membrane binding (loop 2) and toxicity



# Functional Significance of Loops

## Serine proteases

- Enzymes involved in a variety of physiological processes
- Function as digestive enzymes
- Function as regulators & cell differentiation



## Chymotrypsin-like serine protease

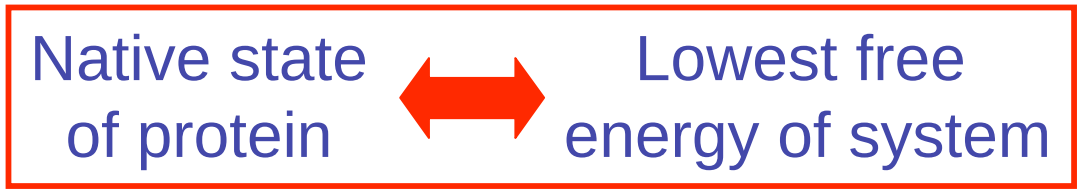
- Catalytic residues bridge beta-barrel
- Substrate specificity determined by three adjacent surface loops
- No direct contact with substrate

## Trypsin-like serine protease

- Subsite specificity in addition to primary
- Two surface loops flank catalytic residues
- Kinetic studies highlight preferential substrate specificity for certain subclasses

# Protein Folding Problem

- To predict a protein's native three-dimensional conformation from its linear amino acid sequence defines the Protein Folding Problem
- Predictive ability would allow for the production of improved drugs, biocatalysts and foster a better understanding of molecular biochemistry and biophysics
- Anfinsen's hypothesis

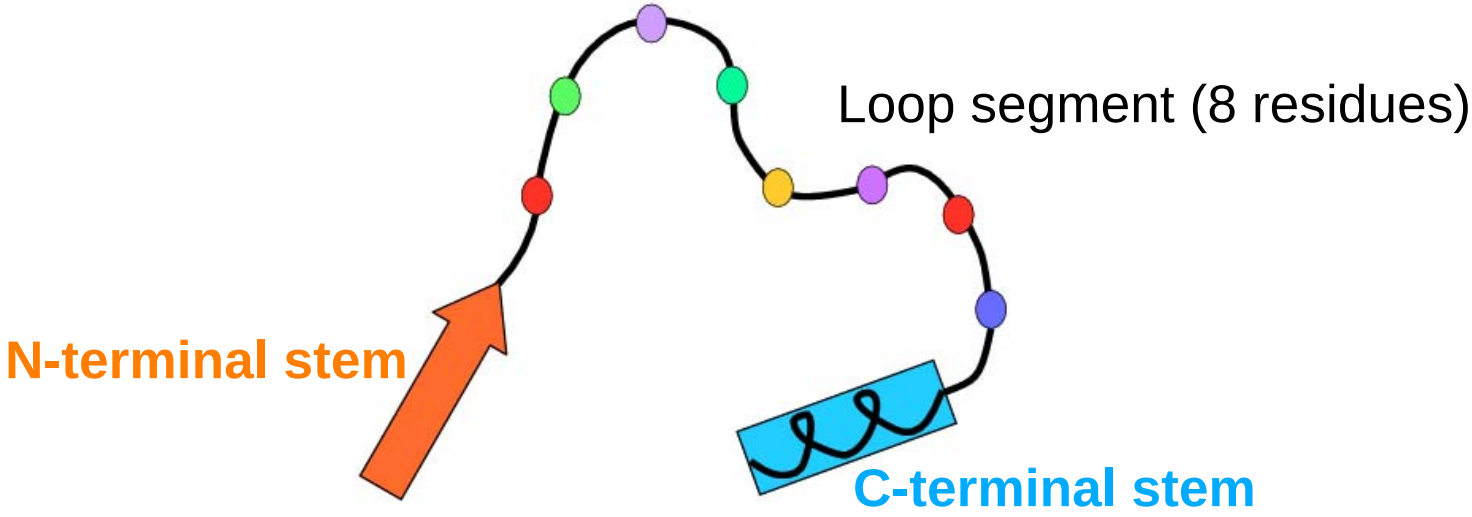



- Express protein free energy as a function of the conformation of the protein (atomic coordinates)
- Use global optimization techniques to deduce the global minimum energy conformation (native state)

# Loop Modeling

## Mini-Protein Folding Problem

- Structure determined from sequence of segment
- Sequence-structure variability is integral to loop functionality
- Loops are not well conserved or regular as with basic secondary structure
- Identical loop segments in different proteins have **unrelated** conformations
- Structure influenced by stem regions that flank loop




## Loop stems

- Stems consist of main chain atoms that precede and follow the loop
- Stem regions offer topological constraints
- Stem regions indicate the orientation for the rest of the protein

# Methods for Loop Modeling

## Database Methods

**Find template that fits the two stem regions of the loop segment**

- Search through all proteins, not just homologous proteins
- Many sequences fit, sort according to
  - Geometric criteria for stem regions (orientation)
  - Sequence similarity between loop and target segments
- Superpose and anneal onto stem regions

Limitations
(1) requires correct loop conformation to exist in database
(2) exponential increase with length for geometric search
(3) only feasible for loops < 8 residues and specific classes

Greer & co-workers

Cohen & co-workers

Levitt

Karplus & coworkers

Chothia & coworkers

## Ab-Initio Methods

**Conformational search guided by scoring or energy function**

- Generate large numbers of loop conformations between loop stems
- Models range from
  - Unified atom models to all atom models (solvation forces)
  - Cartesian to internal coordinates in discrete or continuous space
- Optimization approaches include local minimization, molecular dynamics, systematic searches, simulated annealing, monte carlo

Limitations
(1) native conformation does not provide lowest energy
(2) effectiveness of conformational search procedure
(3) how to handle loop flexibility (conformational entropy)

Sali & co-workers

Scheraga & co-workers

Friesner & co-workers

Karplus & coworkers

Honig & coworkers

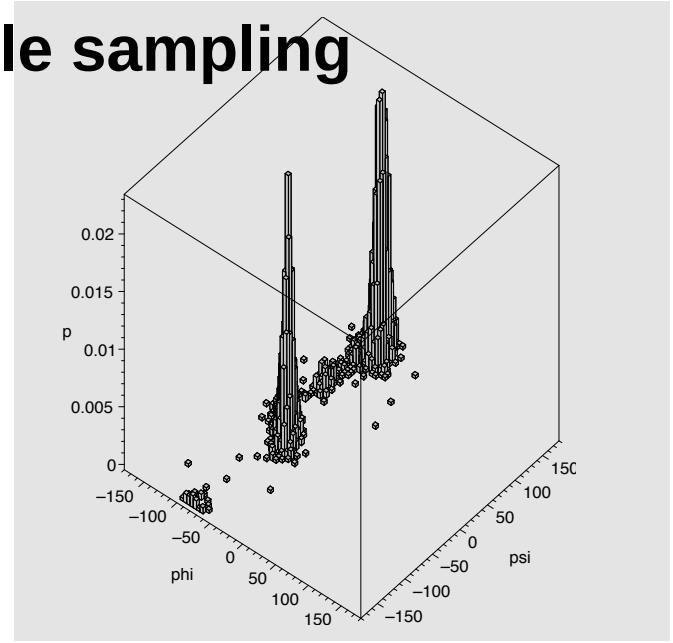
# Existing Approaches to Loop Structure Prediction

## Colony energy minimization

- Xiang, Z., Soto, C., and Honig, B., PNAS 99, 7432-7437, 2002
- Colony energy= potential energy – f(# close neighbors)
- favors conformers in broad energy basins
- random backbone, loop closure,
  2000 conformers for each loop, rotamer libraries for side chains
- 553 loops of lengths 5 to 12 residues
- 2/3 of conformers improved when colony energy is used

## Loop reconstruction with dihedral angle sampling

- DePristo, M., de Bakker, P.I.W., Lovell, S.C., and Blundell, T.L, Proteins 51, 41-55, 2003
- de Bakker, et al., Proteins 51, 21-40, 2003
- backbone angle sampling, probability distributions p(φ,ψ), resolution up to 5°x5°
- 400 loops, 2-12 residues
- energy minimization with AMBER force field, including Generalized Born/surface area solvation model



# Existing Approaches to Loop Structure Prediction

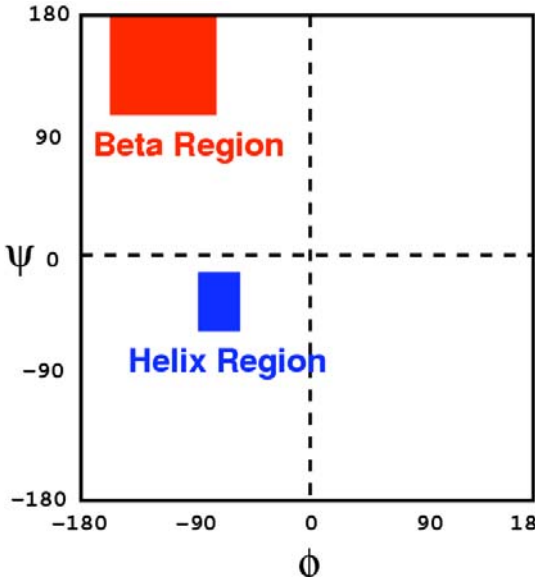
## Loop reconstruction by hierarchical clustering

- Jacobson, M.P., Pincus, D.L., Rappa, C.S., Day, T.J.F., Honig, B., Shaw, D.E., Friesner, R.A., Proteins 55, 351-367, 2004
- Backbone angle sampling with probability functions p(φ,ψ), resolution 5°x5°
- unique in that up to 10^6 conformers are generated
- clustering and filters used to reject decoys before energy minimization
- filters based on information on surrounding protein:
  steric clashes, loop closure, distance to remainder of protein
- test set of 833 loops from 4-12 residues length

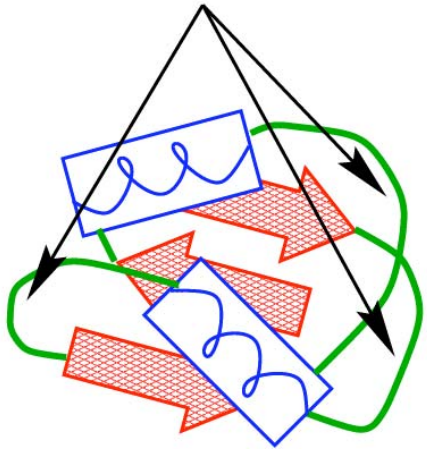
## Data-driven methods

- Baker and coworkers, Proteins 55, 656-677, 2004: fragment database, heuristic scoring function
- Deane and Blundell, Protein Science 10, 599-612, 2001: consensus method that combines database of known loops and set of decoys
- Zhang et al.: efficient statistical energy function that compares favorably to physically based energy functions

# Derivation of Restraints




**Connection between two elements of secondary structure**

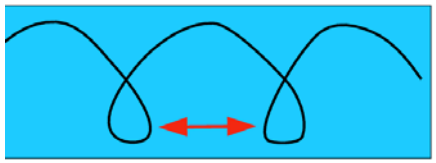


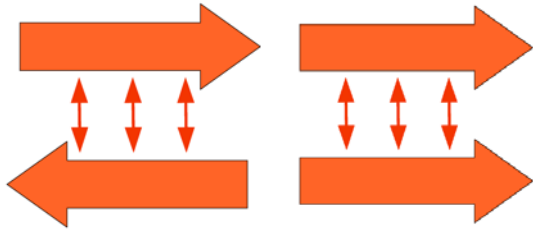
## Dihedral angle restraints

- Backbone dihedral angles restrained according to classification of residue as either helix or strand

## Distance restraints

- Ca-Ca distance restraints for hydrogen bond network of helix (residues i and i+4)
- Ca-Ca distance restraints for hydrogen bonds between residues in opposing strands





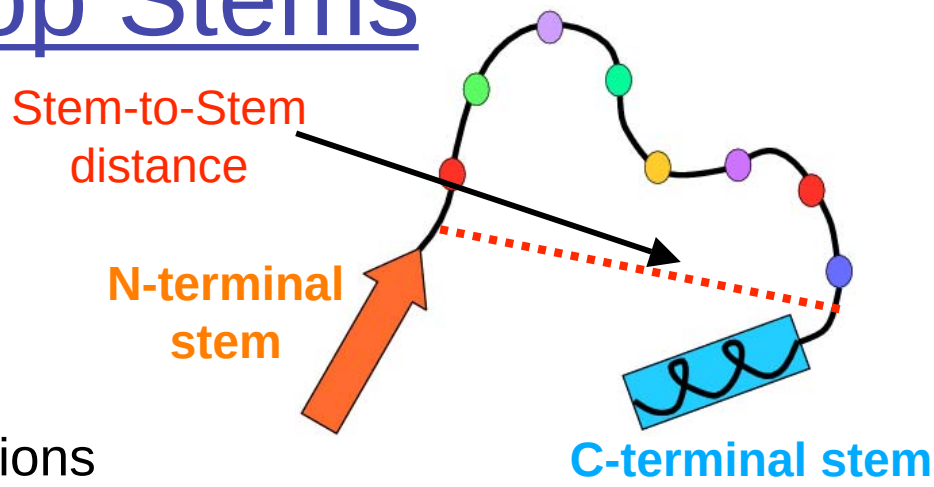
## Bounds on loop residues

- Perform free energy calculations to derive tighter constraints on backbone variables of loop residues

# Using Loop Stems

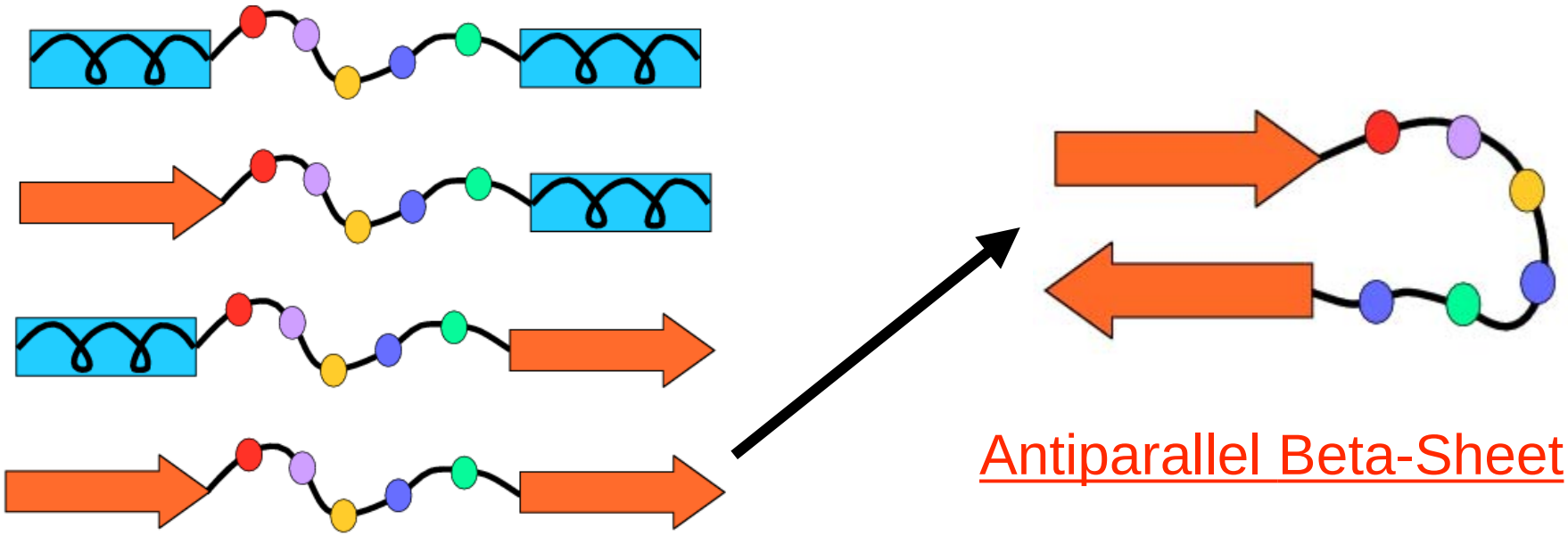
## Topological Constraints

- In general, distance and orientation between loop stems are **not known**




## Available Constraints

- 4 basic classes of loop stem combinations
- Distance constraints only known for **beta-sheet connection**
  - Typically 4.5 - 6.5 angstroms between opposing residues in loop stems
  - Usually such loops are relatively short (< 5 residues)



Antiparallel Beta-Sheet

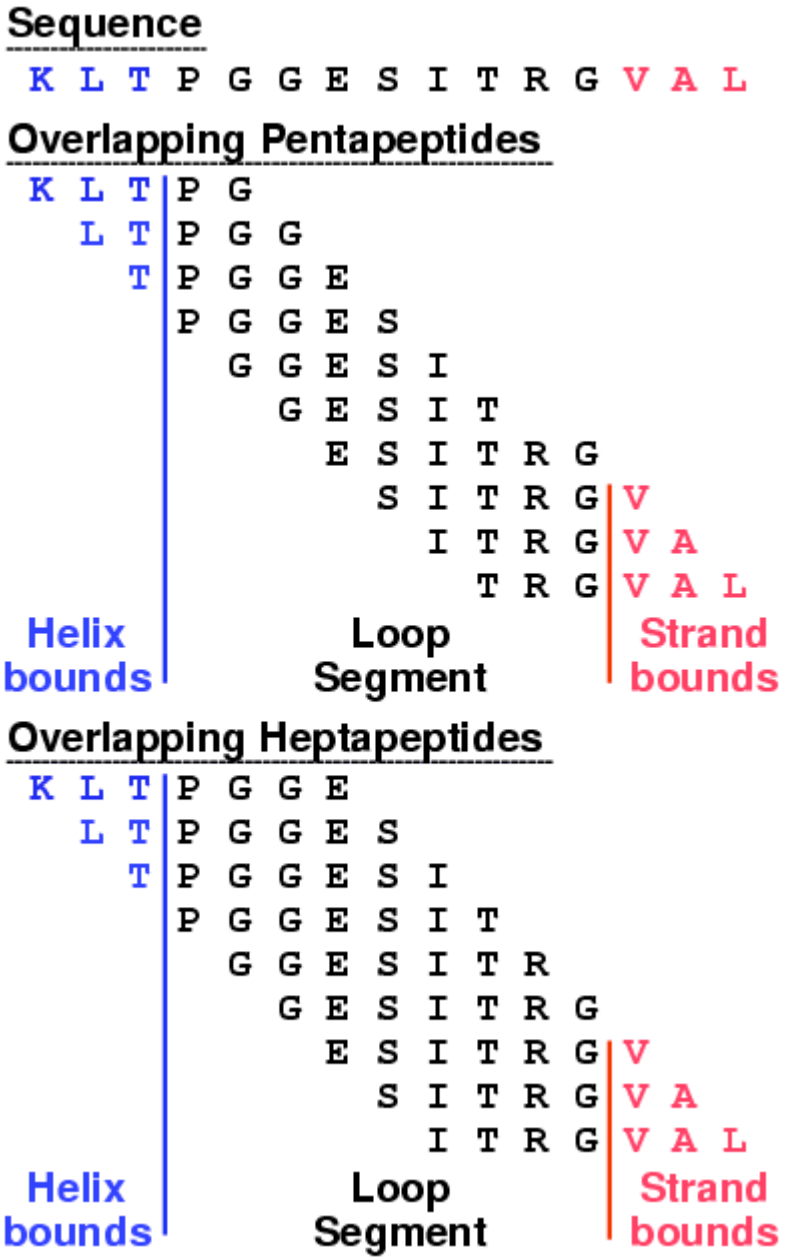
# Protein Loop Prediction Strategies

## Free Energy Calculations for Loop Modeling

- **Overlapping Oligopeptides**
- **Pivot Point Constraints**

# Overlapping Oligopeptides

- Decompose loop segment and **3 flanking stem residues** at each end into smaller oligopeptide sequences
  - Pentapeptides
  - Heptapeptides
  - Nonapeptides
- Impose appropriate bounds on residues in loop stem regions (helix or strand)
- Combine free energy calculations to derive tighter bounds on dihedral angles for residues in loop segment



# Probability of Conformational States

- Calculate probability of conformer *i* from free energy

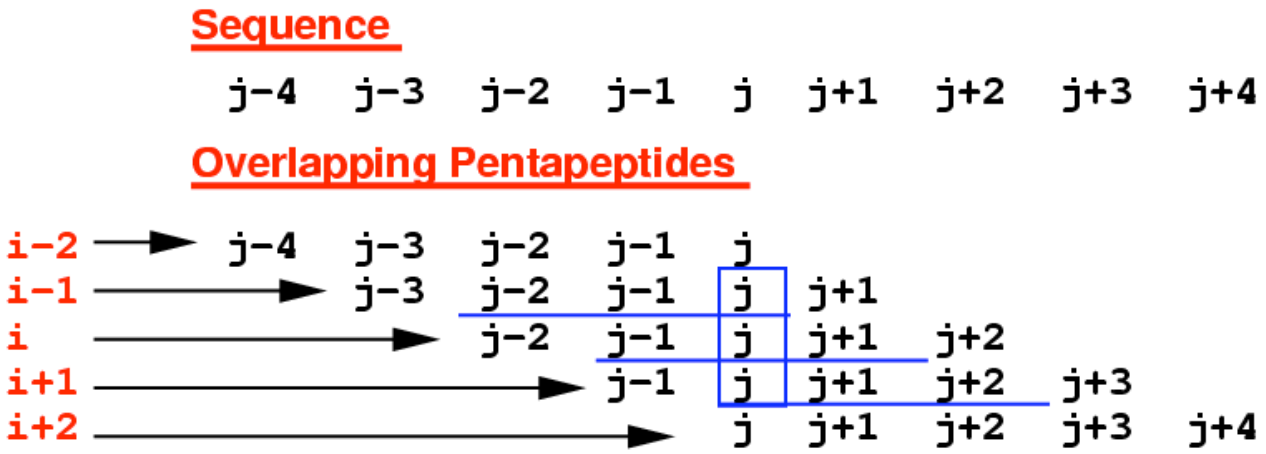
$$p_i = \frac{\exp[-\beta(F_o - F_i)]}{\sum_j \exp[-\beta(F_o - F_j)]}$$

- Cluster probabilities for (YYY) conformational states

$$p_{\text{YYY}} = \sum_{i \in \text{YYY}} p_i$$

- Classify residue of central peptides
- Probability calculation for residue *j* (pentapeptide) is

$$p^j_{\text{YYY}} = \frac{p_{\text{YYY},i-1} + p_{\text{YYY},i} + p_{\text{YYY},i+1}}{3}$$



# Tighter Backbone Constraints

**5 residue loop**
**Helical segment at N-terminal stem**
**Strand segment at C-terminal stem**

Step 1: Initialization

- Select free energy model
- Set bounds for dihedral angles of helix
- Set bounds for dihedral angles of strand

Step 2 : Overlapping Pentapeptides

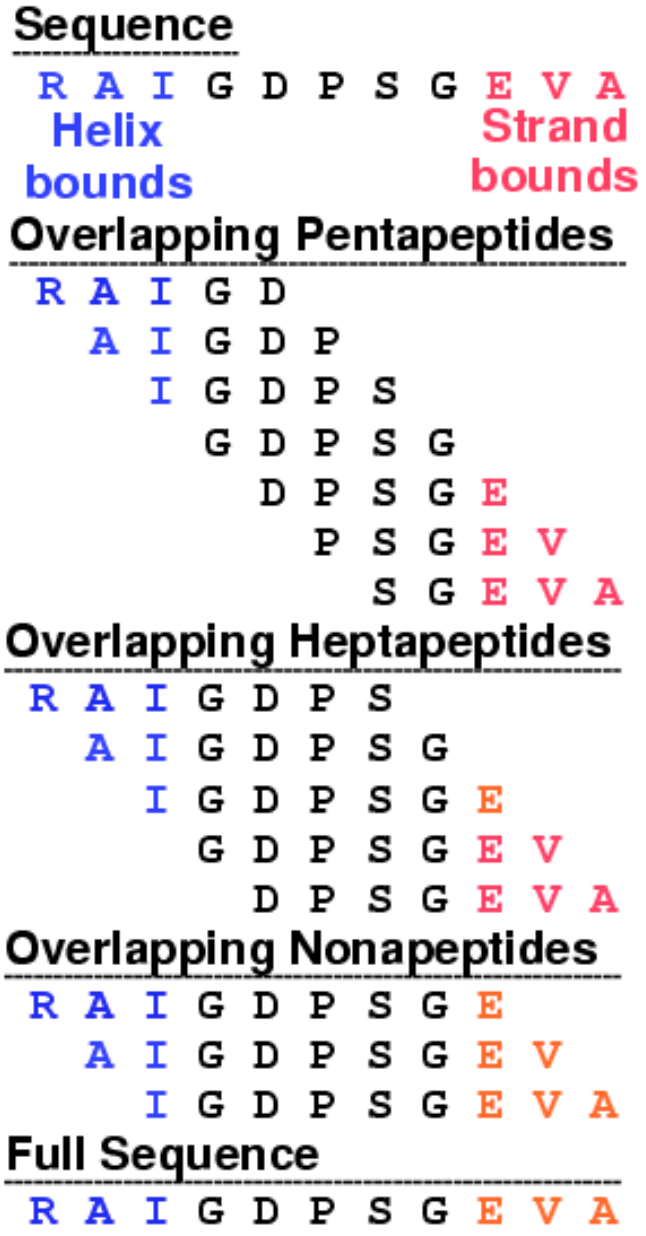
- 7 free energy based optimizations
- Impose appropriate helix/strand bounds
- Calculate cumulative probabilities for conformational state of each residue ($p_{YYY}$)

Step 3 : Overlapping Heptapeptides

- 5 free energy based optimizations
- Impose appropriate helix/strand bounds
- Impose reduced bounds from pentapeptides
- Calculate cumulative probabilities for conformational state of each residue ($p_{YYYYY}$)

Proceed until full sequence simulation

- Use final bounds in tertiary structure prediction



# Overall Free Energy

- **Potential**
  Scheraga & coworkers

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

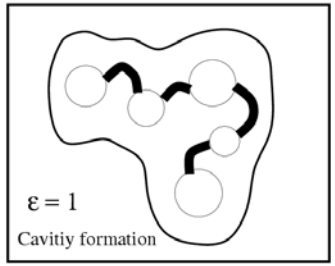
$F_{vac}$ -

- **Entropic**

$$-\frac{k_B}{2}\ln\left[\mathrm{Det}(\mathrm{H}_{\mathrm{vac},\gamma})\right]$$

$TS_{vac}$ +

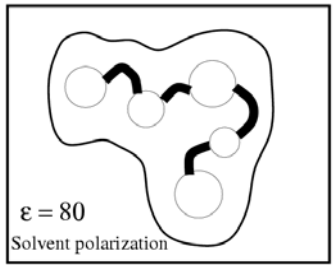
- **Cavity**
  Honig & coworkers
  1988, 1993, 1995




$F_{cavity} = \gamma(SA) + b$

$F_{cavity}$ +

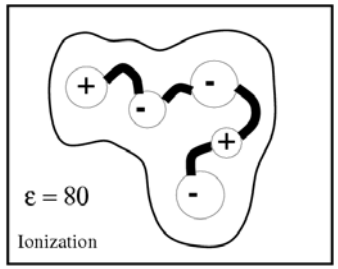
- **Polarization**
  Honig & coworkers
  1988, 1993, 1995




$F_{solv} = F_{polar}(\varepsilon=80) - F_{polar}(\varepsilon=1)$

$F_{solvation}$ +

- **Ionization**
  Honig & coworkers
  1988, 1993, 1995




$F_{ionize}(pH) = kT \ln(Z)$

$F_{ionization}$

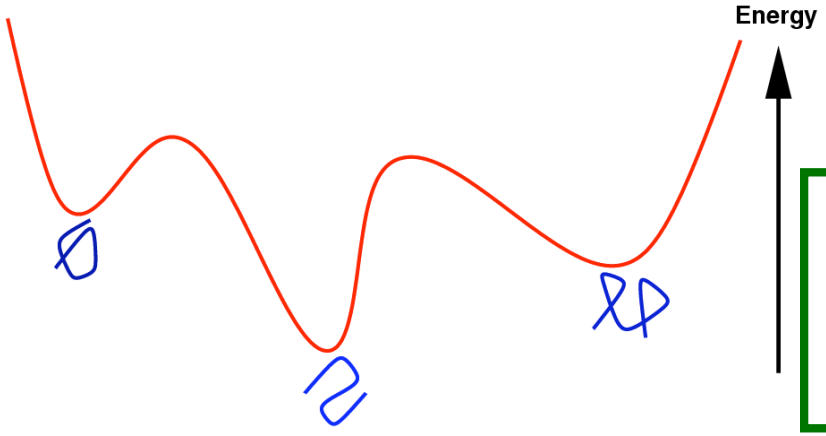
# Ensemble of Low Energy States

Klepeis & Floudas 1999


Generate low energy states along with global minimum energy state

## Mathematical formulation

- Nonconvex optimization problem
- Requires global optimization search




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

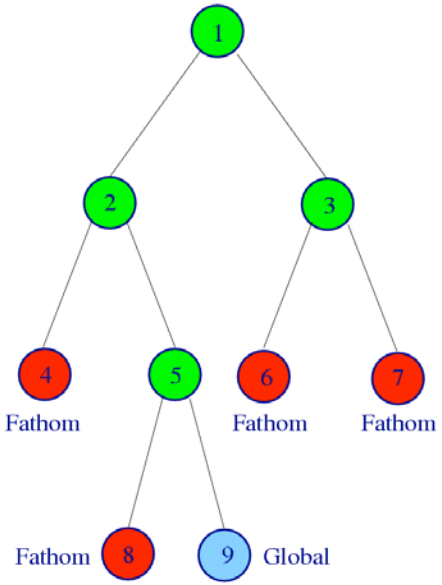
# The αBB Framework

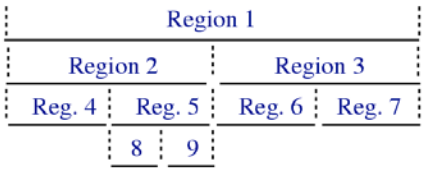
Floudas 2000
Floudas & co-workers
Adjiman et al. 1998,2001

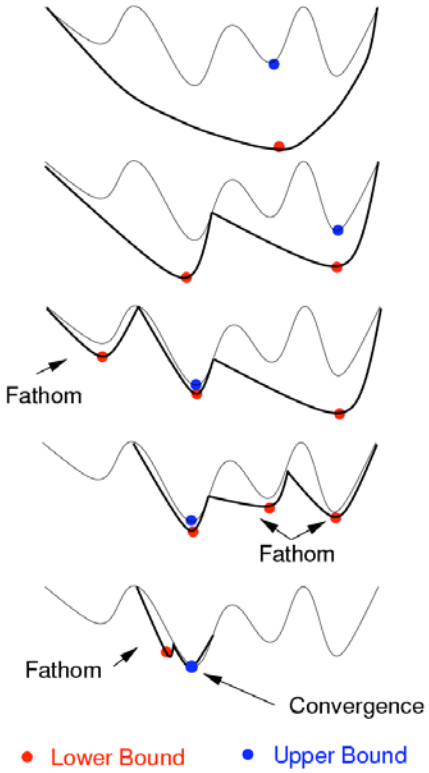
$$\begin{aligned} \min_{\mathbf{x}} \quad & f(\mathbf{x}) \\ \text{s.t.} \quad & \mathbf{h}(\mathbf{x}) = \mathbf{0} \\ & \mathbf{g}(\mathbf{x}) \leq \mathbf{0} \\ & \mathbf{x} \in \mathbf{X} \subseteq \mathcal{R}^n \end{aligned}$$

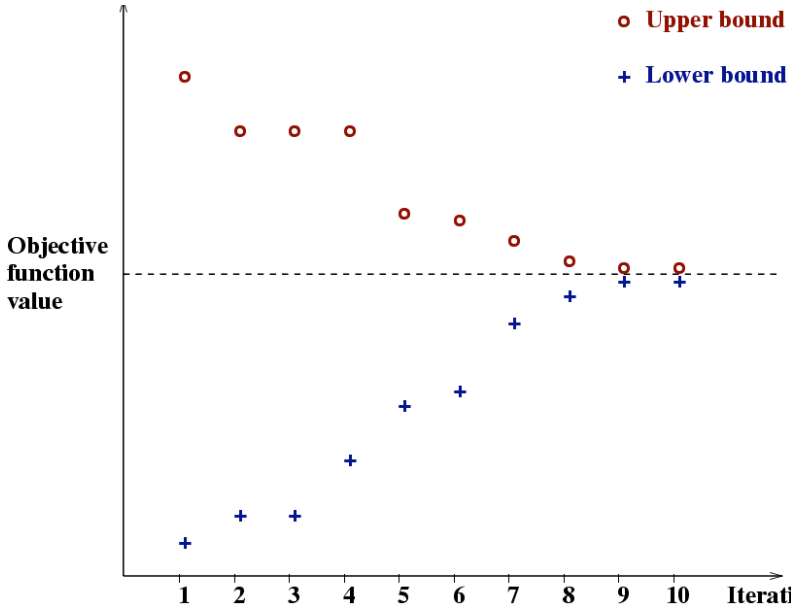
$f$ , $\mathbf{h}$ , $\mathbf{g}$ twice continuously differentiable

- 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



# Protein Loop Prediction Strategies

## Free Energy Calculations for Loop Modeling

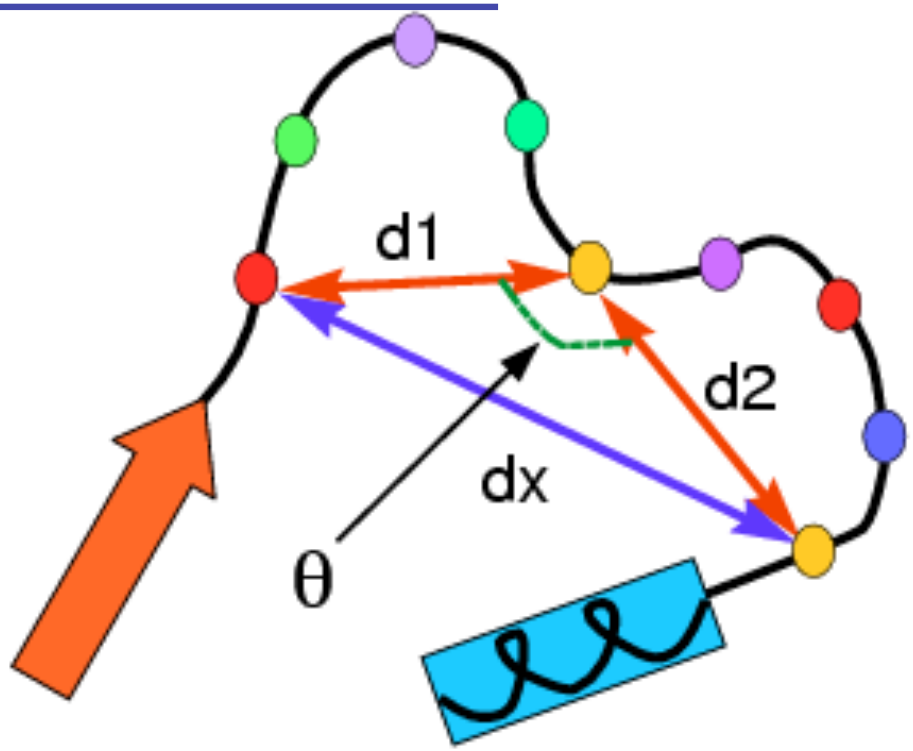
- Overlapping Oligopeptides

- **Pivot Point Constraints**

# Pivot Point Constraints

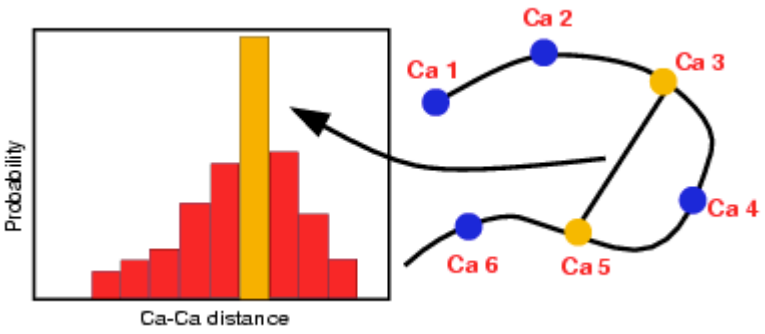
**Goals**

- Improve conformational search
- Derive distance restraints for tertiary structure prediction




**Pivot Point**

- Define Ca of internal loop residue as the pivot point
- From free energy calculations of smaller oligopeptides derive cumulative probability ranges for two Ca to Ca distances between end residues (or other internal residues) and the pivot residue (d1 and d2)
- For most probable distance ranges (d1 and d2) sweep out different ranges of θ and calculate **dx range**
- Impose dx ranges in free energy calculations of larger oligopeptides



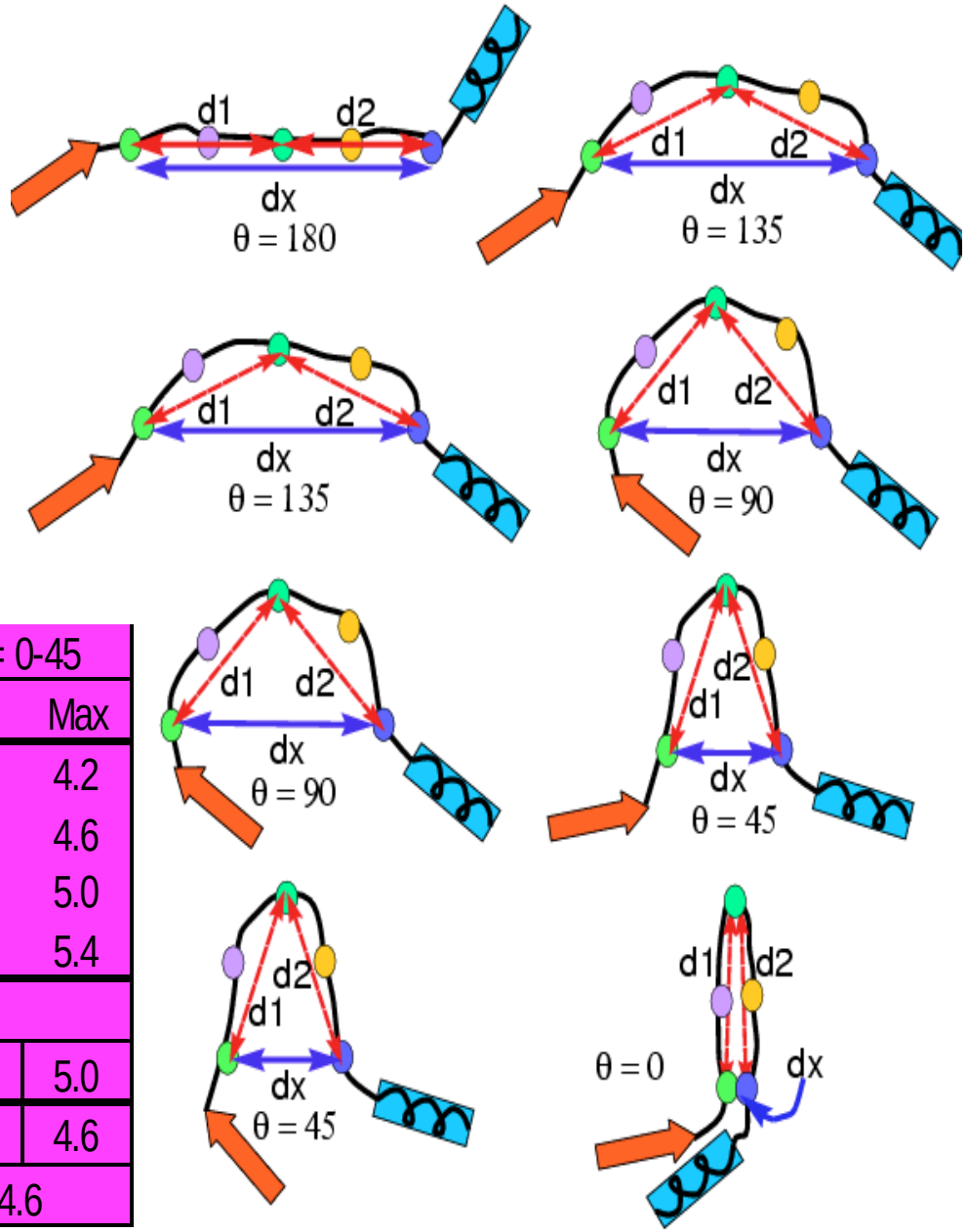
# Tighter Distance Constraints

**5 residue loop**

Pivot Point : Residue 3

d1 : Ca(1) to Ca(3)
d2 : Ca (3) to Ca(5)
**dx** : Ca(1) to Ca(5)

- 4 ranges for **θ** and **dx**
- Free energy calculations for all 4 ranges constrained




| Distance | θ = 135-180 | | θ = 90-135 | | θ = 45-90 | | θ = 0-45 | |
|---|---|---|---|---|---|---|---|---|
| | Min | Max | Min | Max | Min | Max | Min | Max |
| 5.5 | 10.2 | 11.0 | 7.8 | 10.2 | 4.2 | 7.8 | 0.0 | 4.2 |
| 6.0 | 11.1 | 12.0 | 8.5 | 11.1 | 4.6 | 8.5 | 0.0 | 4.6 |
| 6.5 | 12.0 | 13.0 | 9.1 | 12.0 | 5.0 | 9.1 | 0.0 | 5.0 |
| 7.0 | 12.9 | 14.0 | 9.9 | 12.9 | 5.4 | 9.9 | 0.0 | 5.4 |
| | | | | | | | | |
| 6.0-6.5 | 11.1 | 13.0 | 8.5 | 12.0 | 4.6 | 9.2 | 0.0 | 5.0 |
| Unique | 12.0 | 13.0 | 9.1 | 11.1 | 5.0 | 8.5 | 0.0 | 4.6 |
| Range | 1.0 | | 2.0 | | 3.5 | | 4.6 | |

# 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 forcefield

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

## Constraints

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

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

# Torsion Angle Dynamics

Wuthrich & coworkers

Initialization

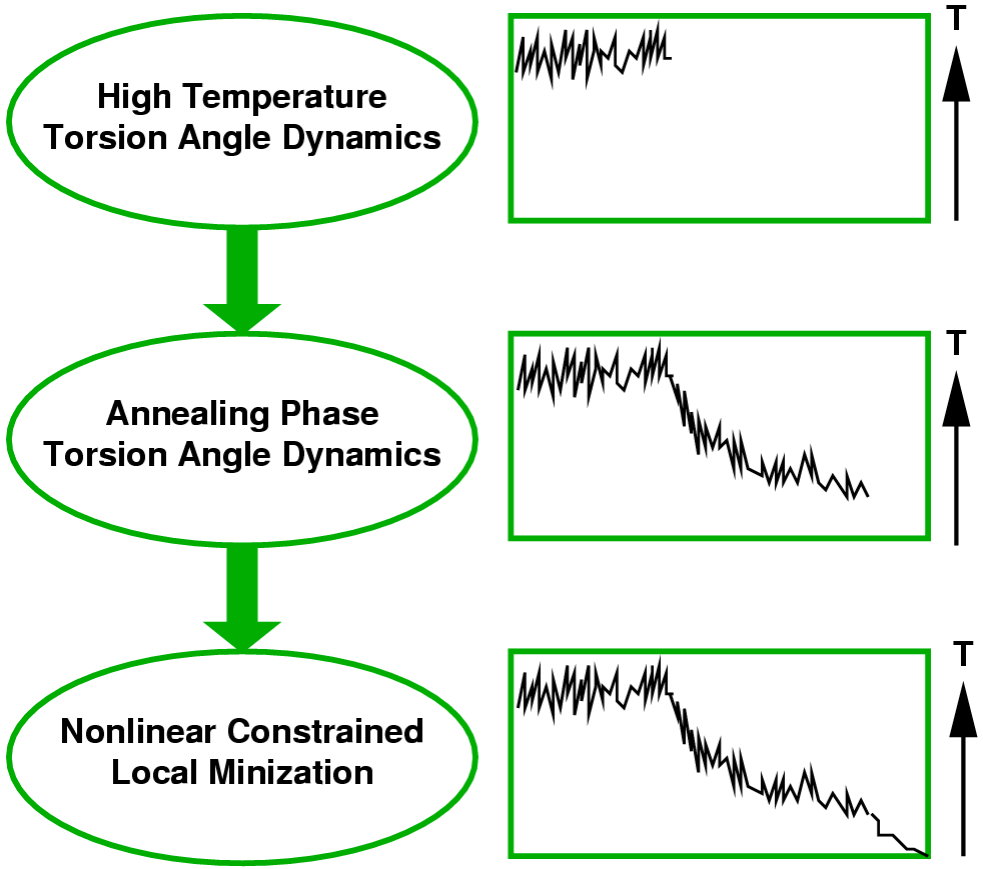
- Difficult to identify low energy feasible structures

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

# Structure Prediction

Bovine Pancreatic Trypsin Inhibitor 56 AAs

Backbone variable restraints

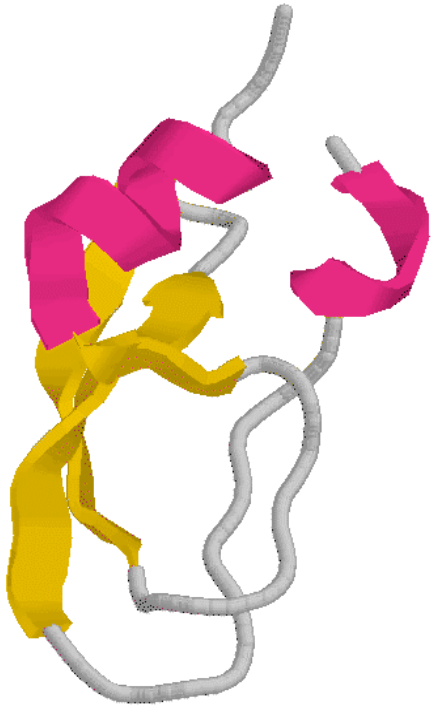
- a-helical : 2-5, 47-54
- β-strand : 17-23, 29-35, 44-46

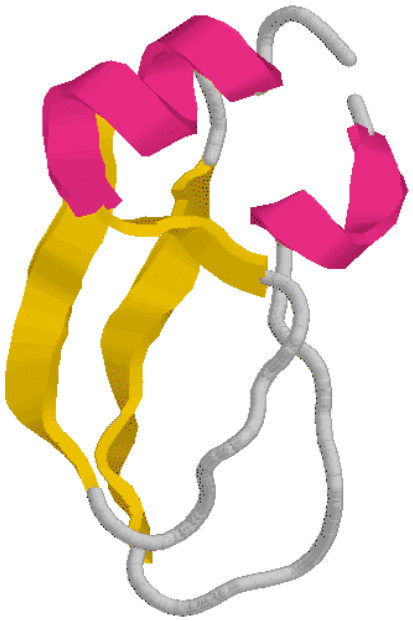
Distance restraints

- Two β-sheet contacts
- 32 lower and upper Ca-Ca for helix and β-sheets
- 6 lower and upper S for disulfide bridge

Tertiary fold

- Best energy : -428.0 kcal/mol
- RMSD : 4.0 A





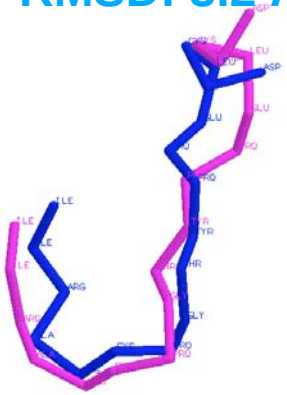
# Bovine Pancreatic Trypsin Inhibitor

Helix 1 | Beta 1 | Beta 2 | Beta 3 | Helix 2

RPDFCLEPPYTGPCKARIIRYFYNAKAGLCQTFVYGGCRAKRNNFKSAEDCMRTCGGA

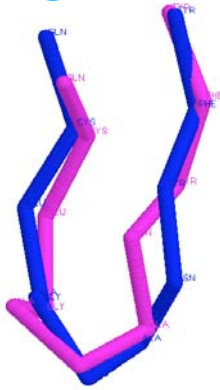
**Loop 1**
**11 residues**
**(Helix-Beta)**
**RMSD: 2.6 A**

**17 residues**
**RMSD: 3.2 A**

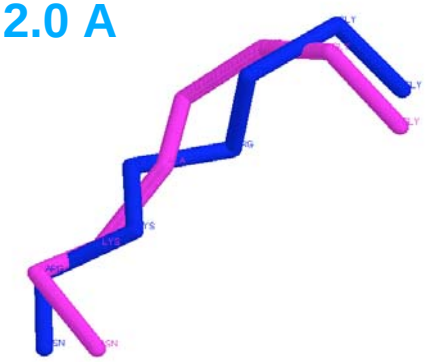


**Loop 2**
**5 residues**
**(Beta-Beta)**
**RMSD: 0.4 A**

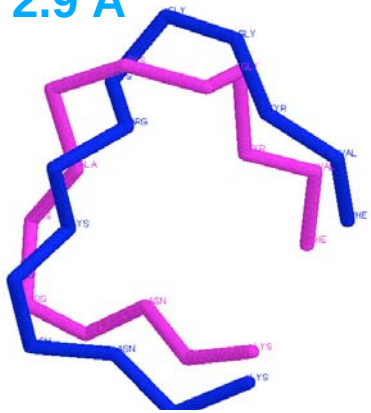
**11 residues**
**RMSD: 1.4 A**



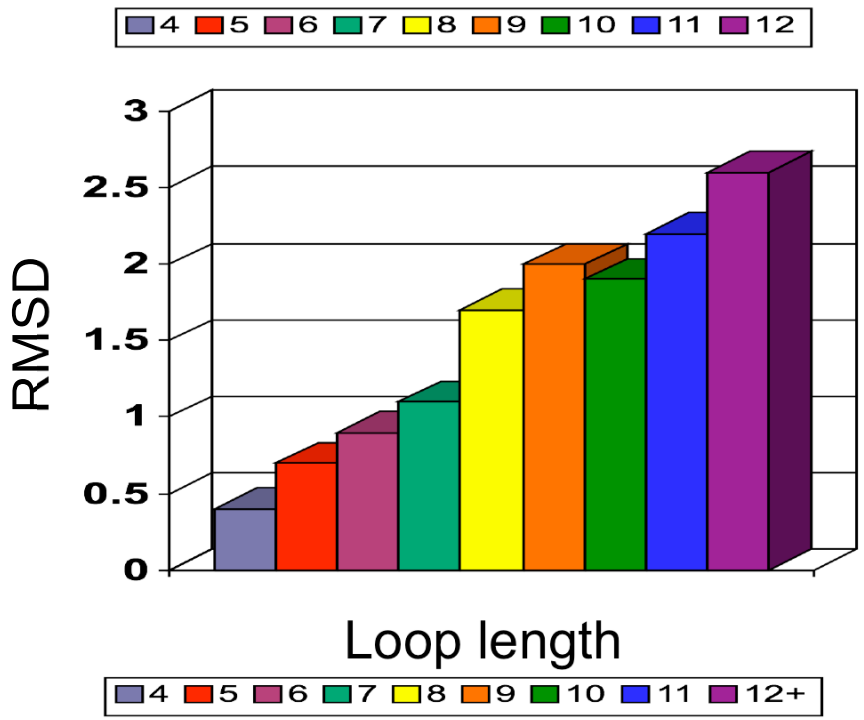
**Loop 3**
**8 residues**
**(Beta-Beta)**
**RMSD: 2.0 A**



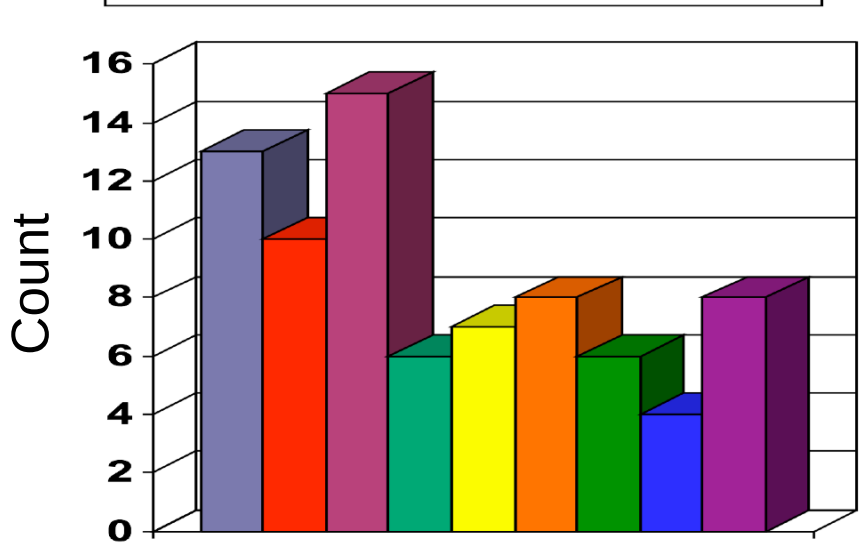
**14 residues**
**RMSD: 2.9 A**



# Large Scale Testing




- Set of 15 proteins
  - Previous CASP competitions
  - Benchmark systems
- Tested within context of ASTRO-FOLD
- Consistent results over all lengths
- Comparable to best results in which loop stems are fixed (+6 residues)



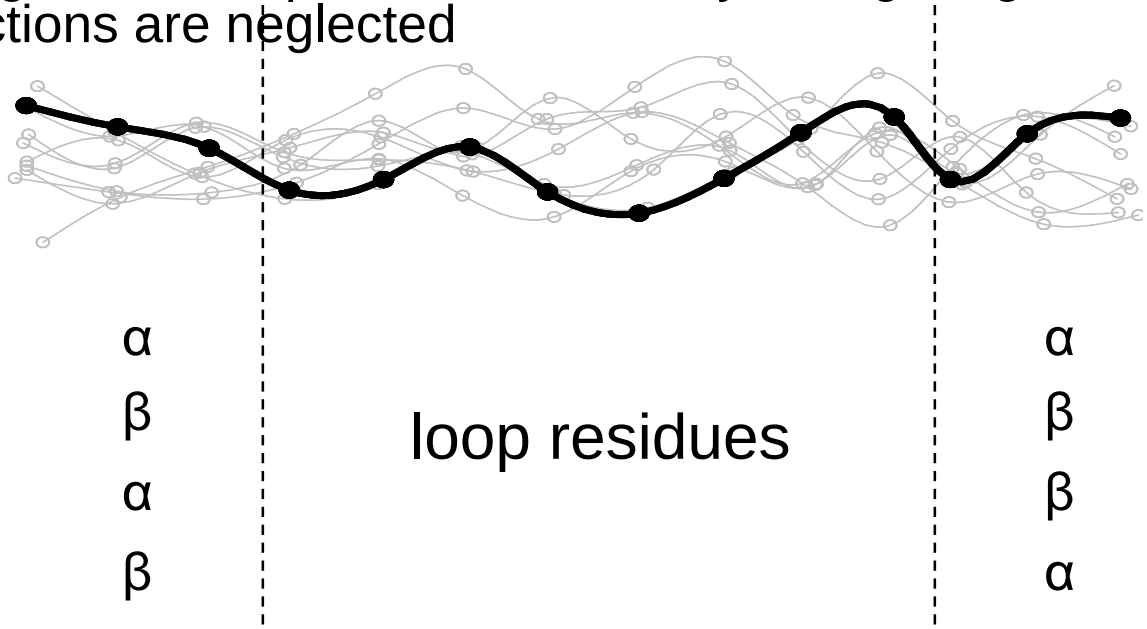

- Set of proteins from CASP5
  - Wide range of loops tested
  - Mixed structure predictions
- Comparisons available starting December 2002

# Loop Structure Prediction with Flexible Stem Geometry

# Flexible Stem Geometry

Assess quality of loop structure prediction with flexible stem geometries

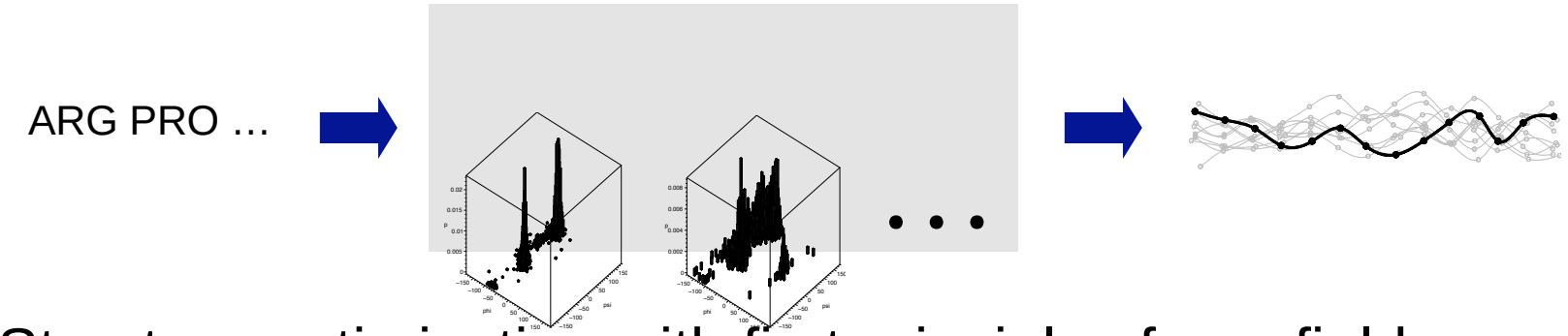
- use in ASTRO-FOLD ab initio structure prediction
  Klepeis, J.L., Floudas, C.A., Biophysical J. 85: 2119-2146, 2003;
  Klepeis, J.L., Floudas, C.A., J. Comput. Chem. 24(2), 191-208, 2003;
  Klepeis, J.L., Floudas, C.A., J. Computat. Chem. 23, 245-266, 2003;
  Klepeis, J.L., Pieja, M.T., Floudas, C.A., Biophysical J. 84, 869-882, 2003;
- investigate limit of prediction accuracy if long range interactions are neglected



# Loop Structure Prediction Methodology

Create ensemble by dihedral angle sampling

- extracted p(φ,ψ) from ~2500 loops
- sampled p(φ,ψ) at 5ºx5º resolution
- created ensembles of 2000 conformers for each loop




Structure optimization with first principles force field

- Dunbrack rotamer library
- ECEPP/3 force field for structure optimization

Clustering to identify conformers that are close to native

# New Use of Clustering

Clustering has been used before to

- group conformers
- select conformers that represent groups

New use of clustering

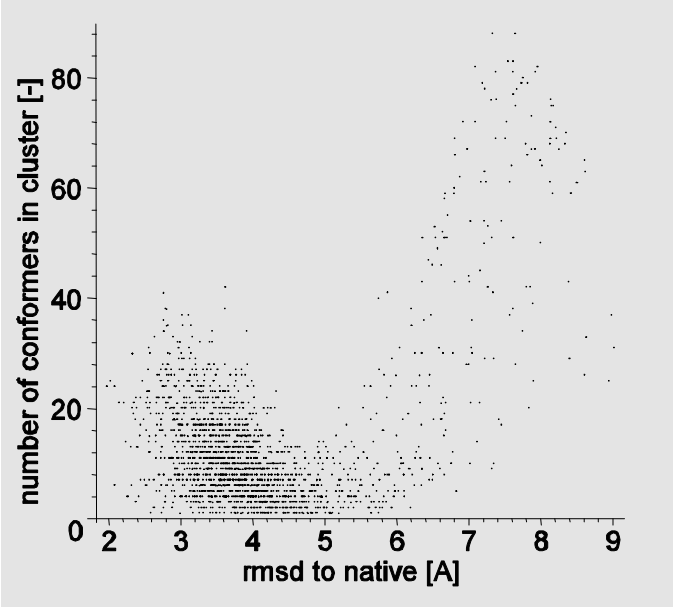
- discard conformers that are far from native

First steps of approach

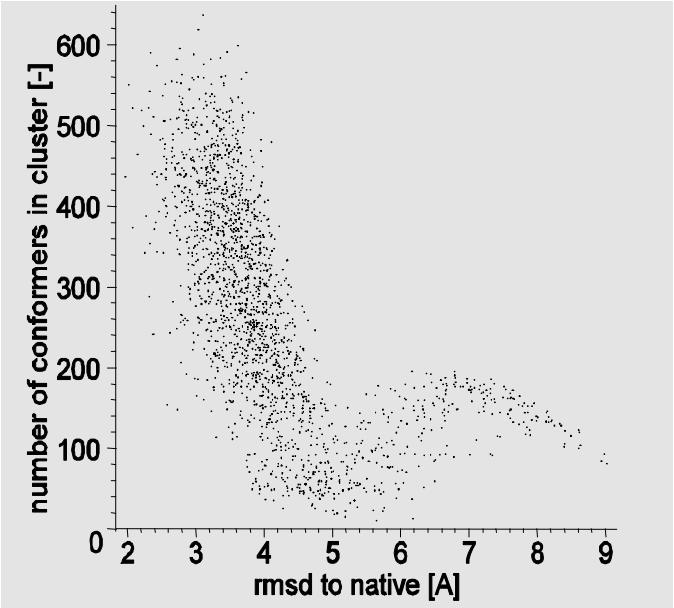
- calculate pairwise RMSDs for ensemble
- choose RMSD threshold *t*
- for each conformer, record number of conformers with RMSD<= *t*

# Clustering Example

- threshold *t*= 3.0Å
- large clusters for small RMSDs unfortunately also for large RMSDs
- not always advisable to consider centroid of largest cluster

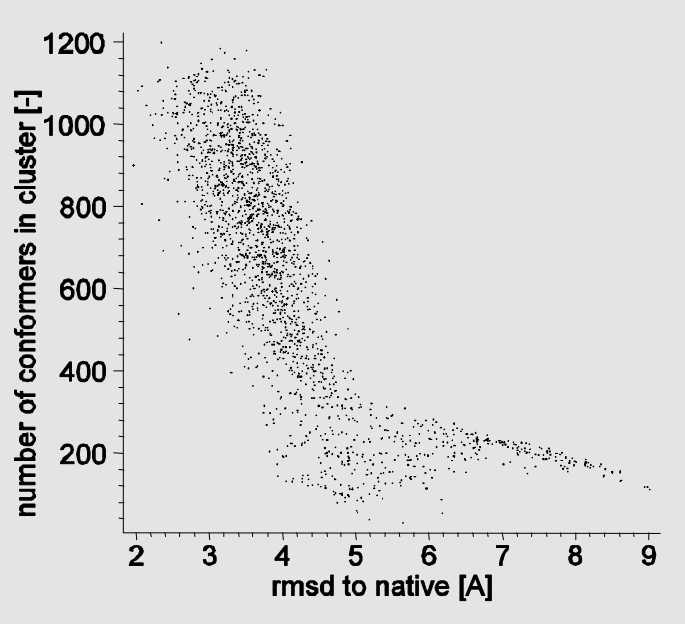



- threshold *t*= 3.5Å
- increasing threshold shows that clusters with large RMSDs are small basins only
- large clusters with small RMSDs survive

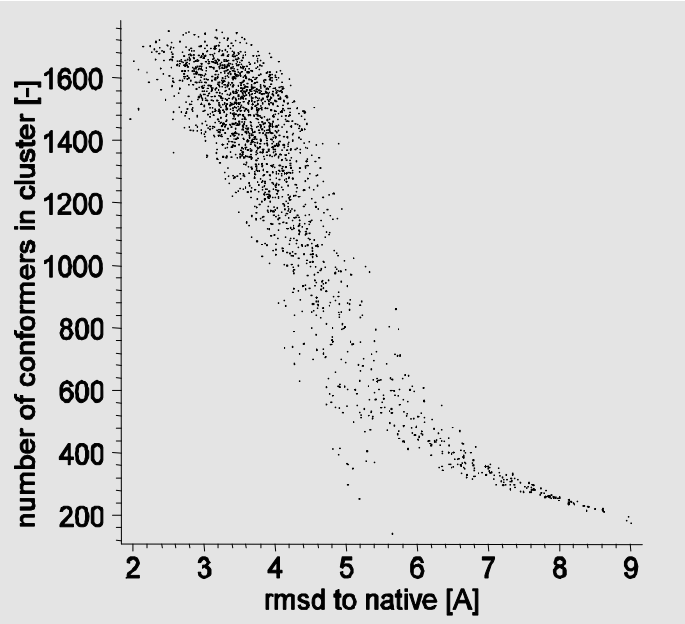


# Clustering Example

- threshold *t*= 4.0Å
- for sufficiently large threshold distribution is monotonous
- tail with large RMSDs becomes apparent




- threshold *t*= 4.5Å
- distribution more conservative the larger threshold
- for sufficiently large threshold clusters of conformers with large RMSDs can be discarded

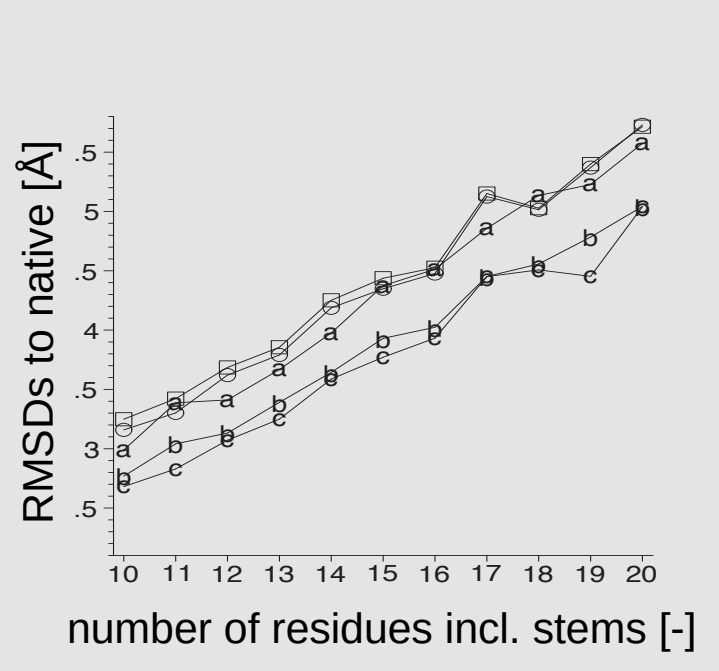


# Iterative Clustering Algorithm

1. Choose thresholds $t$ and $N$,
   choose critical cluster size $N_{crit}$
2. Calculate cluster sizes $N_i$ for all conformers in ensemble
3. If $N_i > N_{crit}$ for all i, stop
4. Discard conformers that generate clusters of size $N_i < N$
5. Go back to step 2

# Results

Treated >3000 loops, length 3+4+3 through 3+14+3



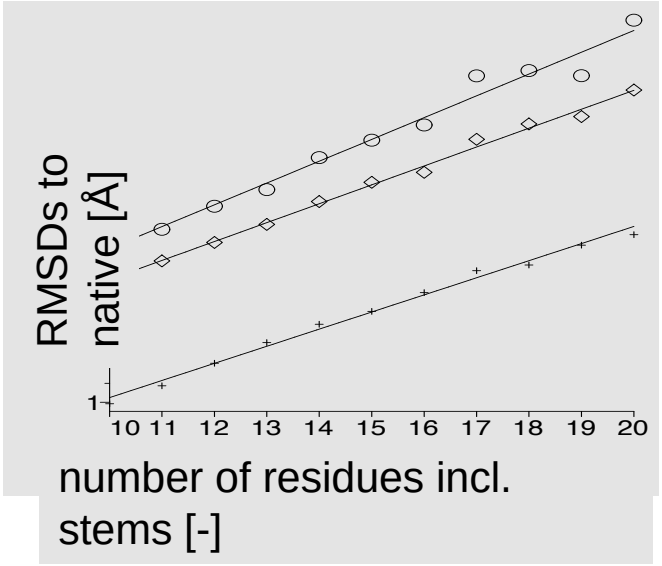

□ energy

○ colony energy

a, b, c cluster size after k=0, 1, 2 clustering steps

- Surprisingly, energy almost as good as colony energy
- Clustering always improves result
- For all loops, algorithm stops after 2 or 3 clustering steps
- RMSD grows only linearly with length for at least 20 residues

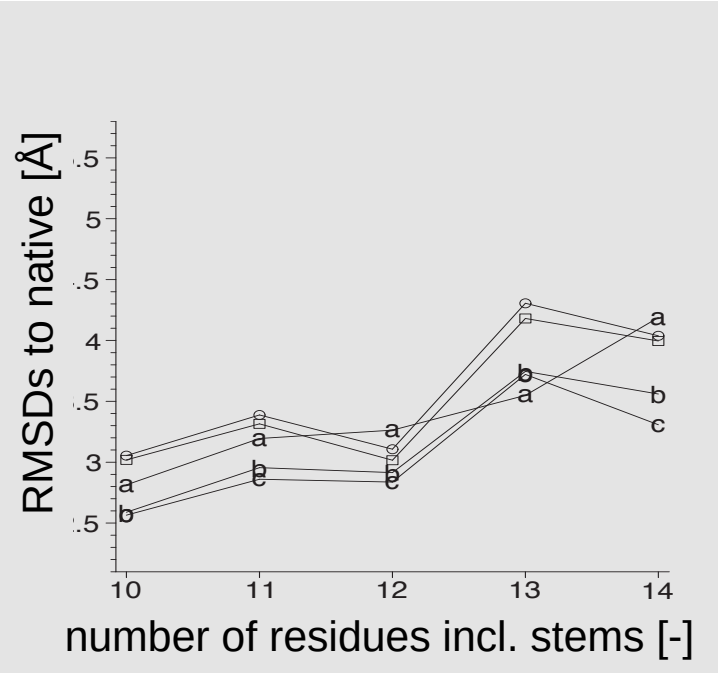
# Quality of Ensembles

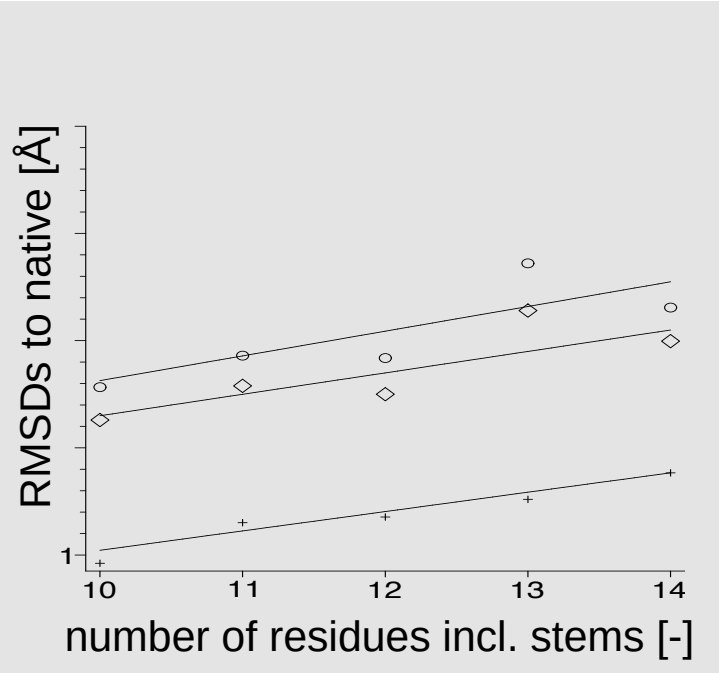



- ○ k= 2 cluster size
- ◇ k= 2 cluster size, select five
- \+ best in ensemble

- Quality of ensembles never restricts result
- Linear for at least up to 20 residues, but slopes differ
- Gap reduced when considering 5 representatives
- Slopes equal when considering 5 representatives

# Results for CASP6 Targets




□ energy

○ colony energy

a, b, c cluster size after k=0, 1, 2 clustering steps

○ k= 2 cluster size

◇ k= 2 cluster size, select five

\+ best in ensemble

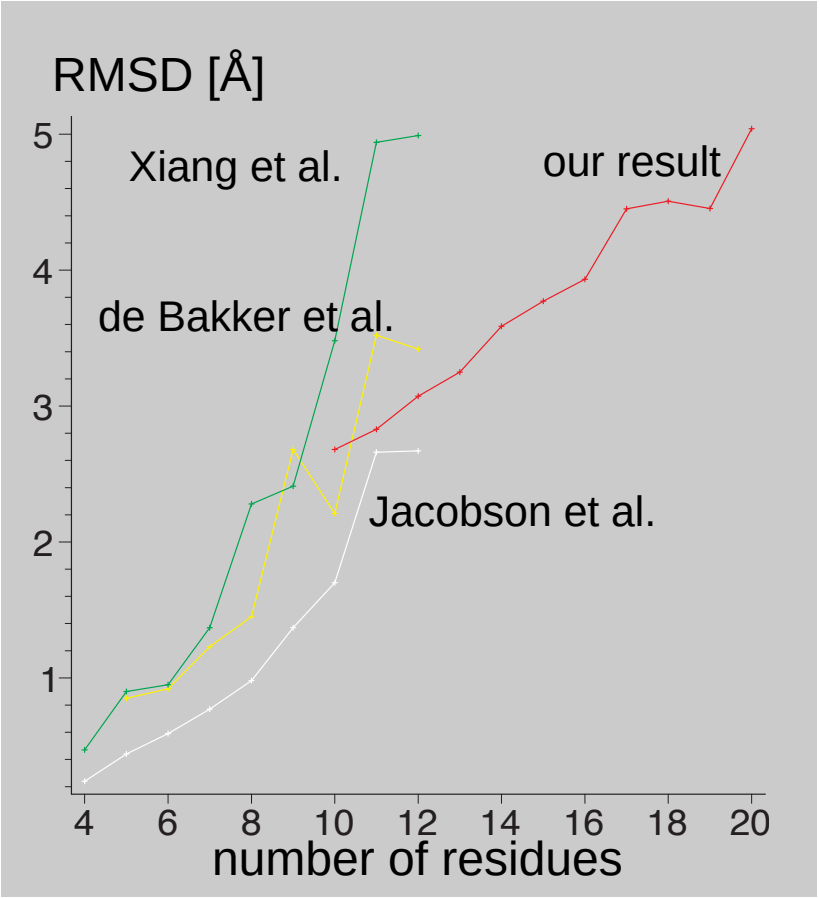
# Comparison to Previous Results

## Comparison difficult

- flexible stem residues
- fixed stems in all previous results

➡ we solve harder problem

- number of residues includes 3+3 stem residues in our case
- stem residues have tighter probability distributions




## Results of comparison

- Jacobson et al. result with fixed stems better

➡ use information on stem geometry, if available

- new method results in very favorable slope
- new method is better than or only slightly worse than methods for fixed stems

# Structure Prediction In Protein Folding: Outline

- Introduction to Protein Structure Prediction
- Free Energy Calculations in Oligo-peptides
- Prediction of Helical Segments
- Prediction of Beta Sheet Topologies
- Prediction of Loop Structures
- Derivation of Restraints
- Prediction of Protein Tertiary Structure

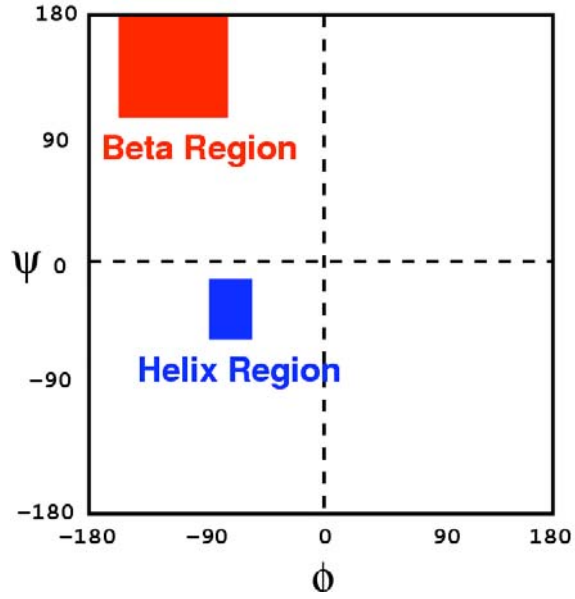
# Derivation of Restraints

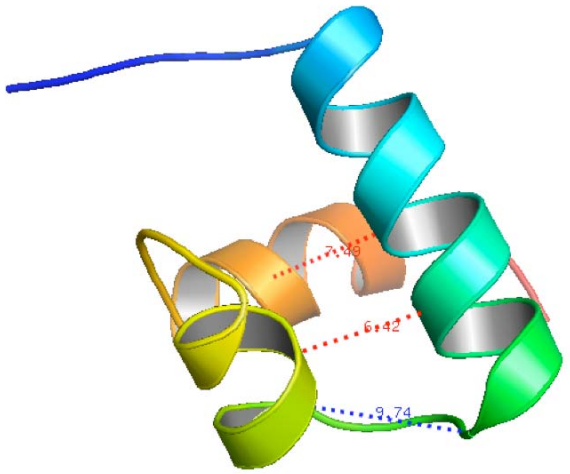
**Relevant References:**


- **Klepeis J.L. and C.A. Floudas, "ASTRO-FOLD: a combinatorial and global optimization framework for ab initio prediction of three-dimensional structures of proteins from the amino acid sequence", Biophysical Journal, 85, 2119-2146 (2003).**
- **McAllister S.R. and C.A. Floudas, ”Enhanced Bounding techniques to Reduce the Protein Conformational Search Space”, submitted for publication, 2008.**

# Derivation of Restraints

- Dihedral angle restraints
  - For residues with $\alpha$-helix or $\beta$-sheet classification
  - For loop residues using the best identified conformer from loop modeling efforts
- Distance restraints
  - Helical hydrogen bond network (i,i+4)
  - $\alpha$-helical topology predictions
  - $\beta$-sheet topology predictions



Klepeis, JL and Floudas, CA. Journal of Global Optimization. (2003)

# Structure Prediction In Protein Folding: Outline

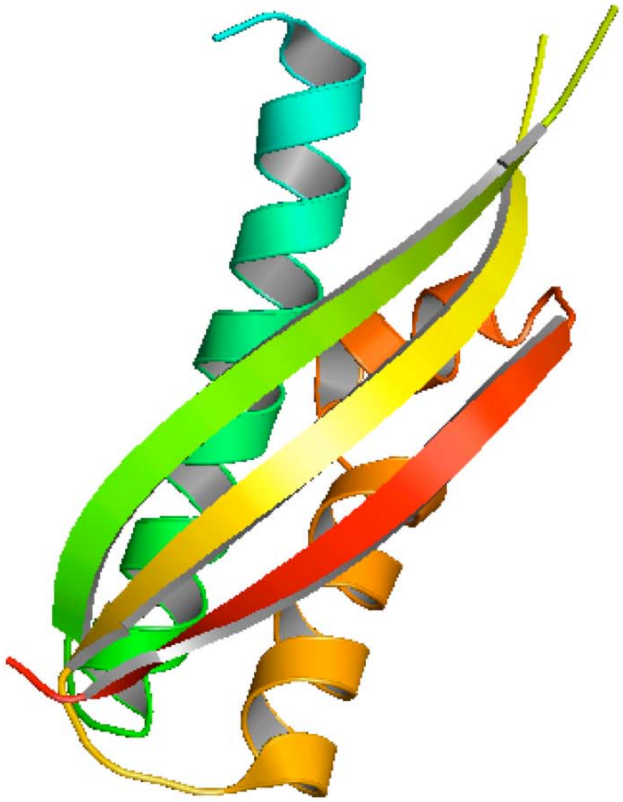
- Introduction to Protein Structure Prediction
- Free Energy Calculations in Oligo-peptides
- Prediction of Helical Segments
- Prediction of Beta Sheet Topologies
- Prediction of Loop Structures
- Derivation of Restraints
- Prediction of Protein Tertiary Structure

# Prediction of Protein Tertiary Structure

## Relevant References:

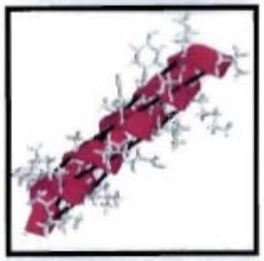

- **Klepeis J.L. and C.A. Floudas, "Ab Initio Tertiary Structure Prediction of Proteins", Journal of Global Optimization, 25, 113-140 (2003).**
- **Klepeis J.L., M. Pieja, and C.A. Floudas, "A New Class of Hybrid Global Optimization Algorithms for Peptide Structure Prediction: Integrated Hybrids", Computer Physics Communications, 151, 121-140 (2003).**
- **Klepeis J.L., M. Pieja, and C.A. Floudas, "A New Class of Hybrid Global Optimization Algorithms for Peptide Structure Prediction: Alternating Hybrids and Application fo Met-Enkephalin and Melittin", Biophysical Journal, 84, 869-882 (2003).**
- **Klepeis J.L. and C.A. Floudas, "ASTRO-FOLD: a combinatorial and global optimization framework for ab initio prediction of three-dimensional structures of proteins from the amino acid sequence", Biophysical Journal, 85, 2119-2146 (2003).**
- **Klepeis J.L., Y. Wei, M.H. Hecht, and C.A. Floudas, "Ab Initio Prediction of the 3-Dimensional Structure of a De Novo Designed Protein: A Double Blind Case Study", Proteins, 58, 560-570 (2005).**

# Tertiary structure prediction



# ASTRO-FOLD

(Klepeis & Floudas, 2002c)



## Helix Prediction

- Detailed Modeling
- Simulations of Local Interactions

(Free Energy Calculations)

## Beta Sheet Prediction

- Novel Modeling of Beta Sheet Formation
- Predict List of Optimal Arrangements

(Combinatorial Optimization)

## Derivation of Restraints

## Overall 3D Structure Prediction

- Structural Data from Previous Stages
- Prediction via Novel Solution Approach

(Global Optimization & Molecular Dynamics)

# Tertiary Structure Prediction : Key Ideas

(Klepeis & Floudas, 2002c)

- **Utilization of Helix and Beta Predictions**

  Enforce predictions of secondary structure and $\beta$ sheet configuration through rigorous constraints

- **Mathematical Formulation**

  Formulate tertiary structure prediction problem as a constrained global optimization problem

- **Energy Modeling**

  Model proteins using detailed atomistic level force-field

- **Global optimization approach**

  (Floudas & coworkers, 1999,2000)

  Predict overall tertiary structure using combination of $\alpha$BB global optimization (Floudas & coworkers, 1994,1995,1996,1998,1999) and torsion angle dynamics (Jain & coworkers, 1993)

# 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 forcefield (Scheraga & coworkers)

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

## Constraints

- Enforce bounds on backbone variables $[\boldsymbol{\phi}^L, \boldsymbol{\phi}^U]$
- Enforce upper and lower distances bounds through $N_{\text{con}}$ constraints with form of square well potential

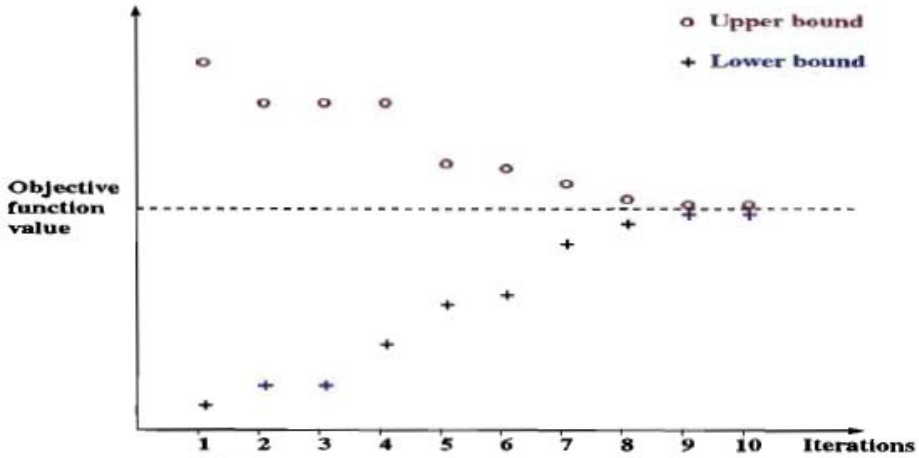
$$E_{\text{distance}} = \sum_{j \in \text{upper}} \begin{cases} A_j (d_j - d_j^{\text{U}})^2 & \text{if } d_j > d_j^{\text{U}} \\ 0 & \text{otherwise} \end{cases} + \sum_{j \in \text{lower}} \begin{cases} A_j (d_j - d_j^{\text{L}})^2 & \text{if } d_j < d_j^{\text{L}} \\ 0 & \text{otherwise} \end{cases}$$

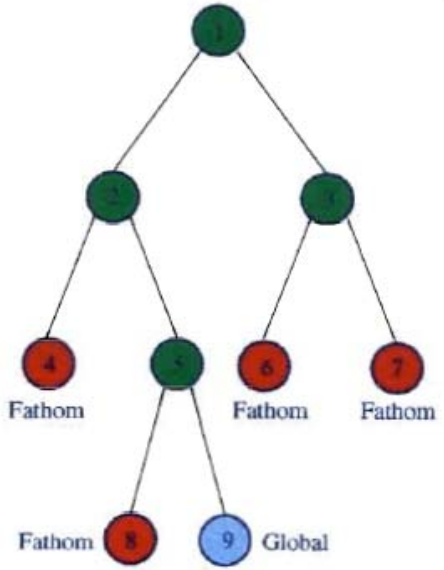
# The $\alpha$BB Framework

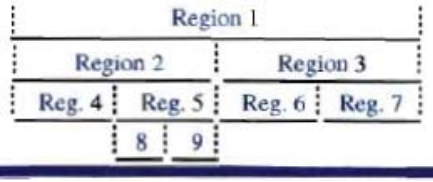
$$
\begin{array}{lll}
\min\limits_{\mathbf{x}} & f(\mathbf{x}) & \\
\text{s.t.} & \mathbf{h}(\mathbf{x}) = \mathbf{0} & \\
& \mathbf{g}(\mathbf{x}) \leq \mathbf{0} & \\
& \mathbf{x} \in \mathbf{X} \subseteq \mathcal{R}^n &
\end{array}
$$

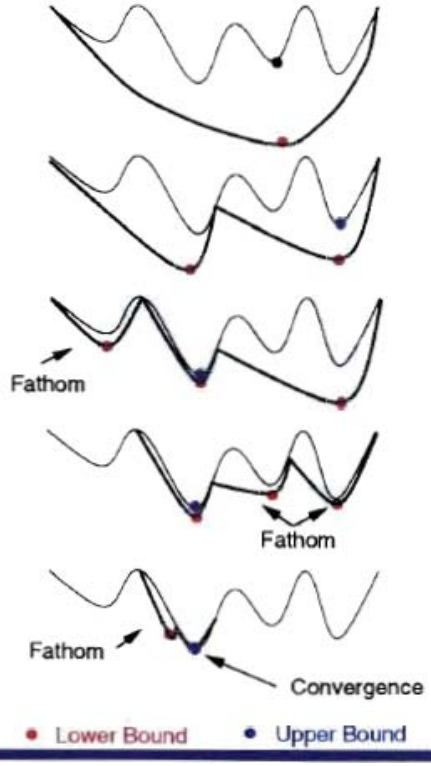
$f$ , $\mathbf{h}$ , $\mathbf{g}$ twice continuously differentiable

- 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 branch and bound tree
- Guaranteed $\epsilon$-convergence for $C^2$ NLPs



# Torsion Angle Dynamics

Difficult to identify low energy structures satisfying constraints because large protein systems possess

- High dimensionality
- Sparse sets of restraints

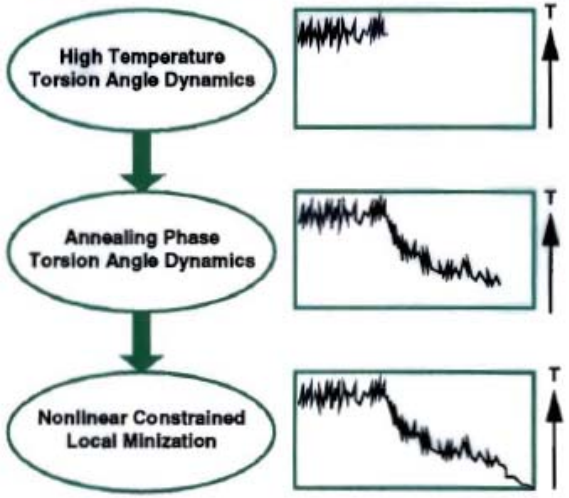
**Torsion Angle Dynamics (TAD)**

- Rapidly identify feasible low energy structures
- Fast evaluation of simplified force field
- Unconstrained formulation using penalty functions

**Implementation**

- Solve equations of motion as preprocessing for each constrained minimization in $\alpha$BB approach

$$\mathcal{M}(\boldsymbol{\theta})\ddot{\boldsymbol{\theta}} + \mathcal{C}(\boldsymbol{\theta}, \dot{\boldsymbol{\theta}}) = 0$$



# Structure Prediction : BPTI

## Backbone variable restraints

- $\alpha$-helical : 2-5, 47-54
- $\beta$-strand : 17-23, 29-35, 44-46

## Distance restraints

- 2 $\beta$ sheet contacts
- 32 lower and upper $C^{\alpha}$-$C^{\alpha}$ for helix and $\beta$-sheets
- 6 lower and upper $S^{\gamma}$ for disulfide bridges

## Tertiary Fold

- Best energy : -428.0 kcal/mol
- RMSD : 4.0 $\AA$

# Structure Prediction

T0114 87 AAs

Backbone variable restraints

- β-strand : 12-17, 22-27, 31-37, 39-43, 48-54, 61-67, 78-86

Distance restraints

- 5 β-sheet contacts
- 68 lower and upper Ca-Ca for helix and β-sheets
- 2 lower and upper S for disulfide bridge

Tertiary fold

- Best energy : -530.0 kcal/mol
- RMSD : 4.6 A

# Structure Prediction : T0114

## Backbone variable restraints

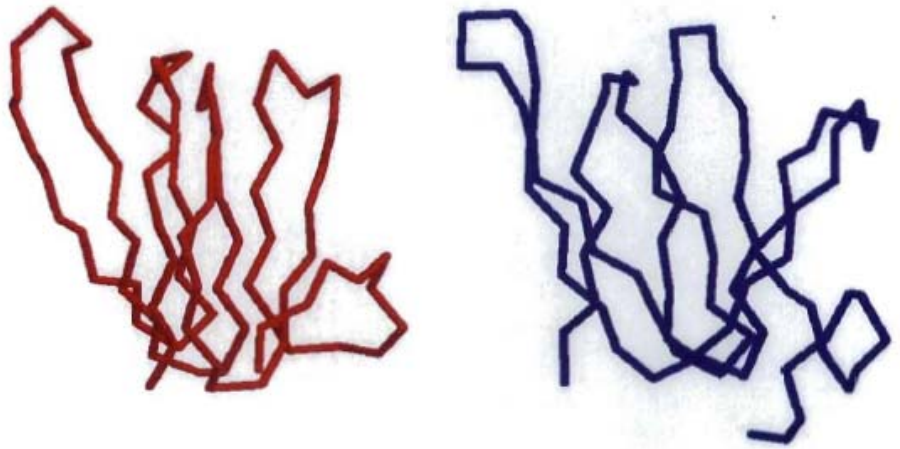
- $\beta$-strand : 12-17, 22-27, 31-37, 39-43, 48-54, 61-67, 78-86

## Distance restraints

- 5 $\beta$ sheet contacts
- 68 lower and upper $C^{\alpha}$-$C^{\alpha}$ for $\beta$-sheets
- 2 lower and upper $S^{\gamma}$ for disulfide bridges

## Tertiary Fold

- Best energy : -530.0 kcal/mol
- RMSD : 4.6 $\AA$



# CASP5 Structure Prediction

T0176 100 AAs

Backbone variable restraints

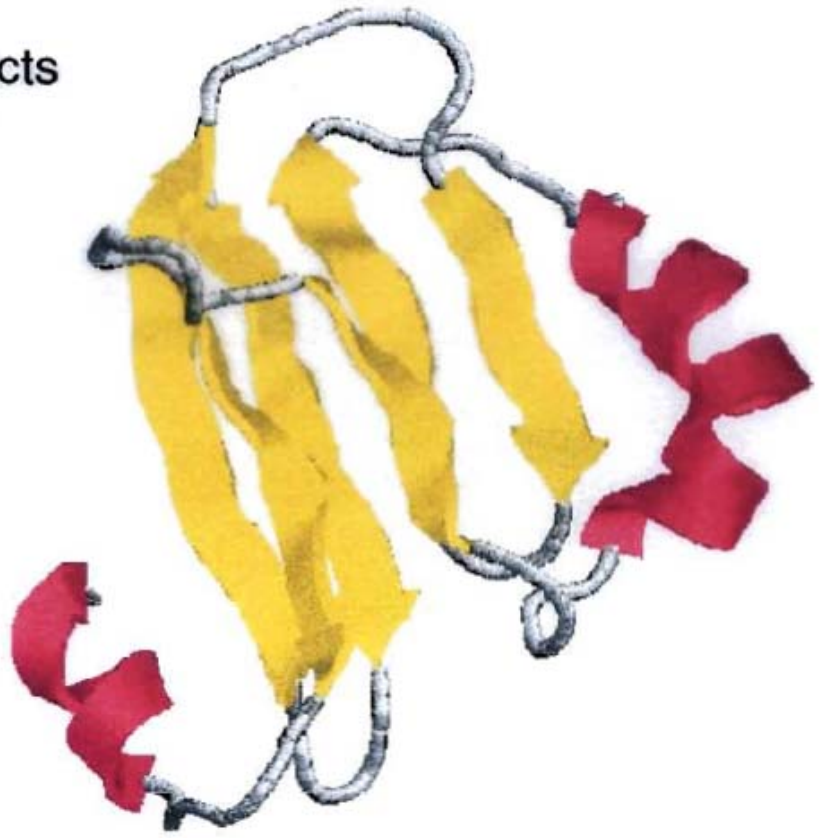
- α-helix : 51-63, 94-99
- β-strand : 4-10, 15-22, 30-34, 38-43, 69-73, 82-87

Distance restraints

- 5 antiparallel β-sheet contacts
- 38 lower and upper Ca-Ca for helix and β-sheets

Tertiary fold

- Best energy : -423.0 kcal/mol



# CASP5 Structure Prediction

**T0188** 124 AAs

Backbone variable restraints

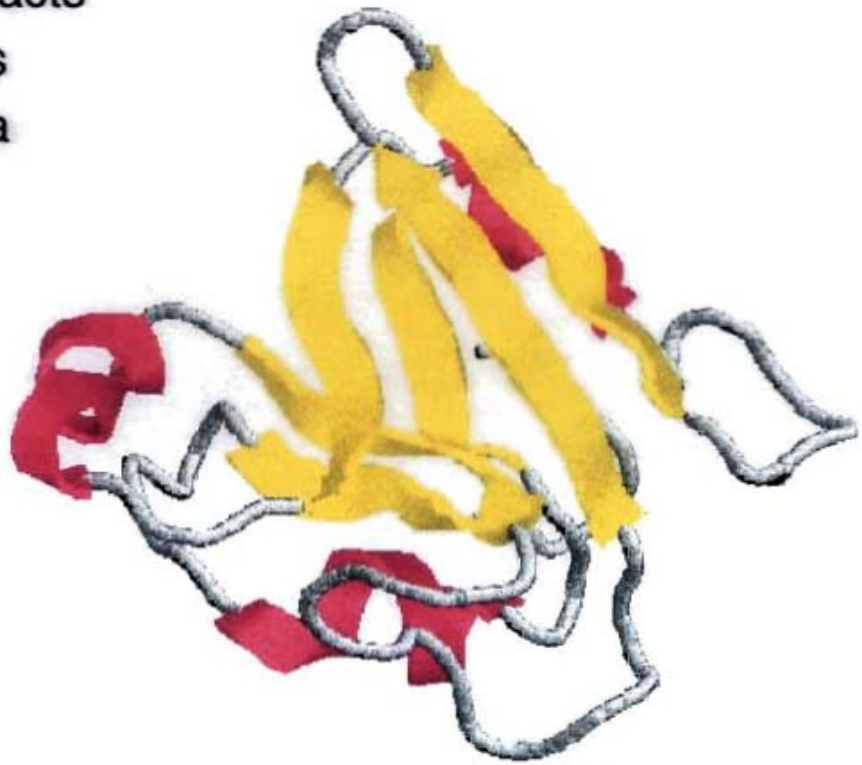
- α-helix : 57-62, 76-83, 96-103
- β-strand : 1-7, 26-32, 37-43, 66-71, 86-90, 113-116

Distance restraints

- 2 antiparallel β-sheet contacts
- 3 parallel β-sheet contacts
- 31 lower and upper Ca-Ca for helix and β-sheets

Tertiary fold

- Best energy : -484.0 kcal/mol



# Constrained optimization

- Problem definition

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

- Atomistic level force field (ECEPP/3)

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

- Distance constraints

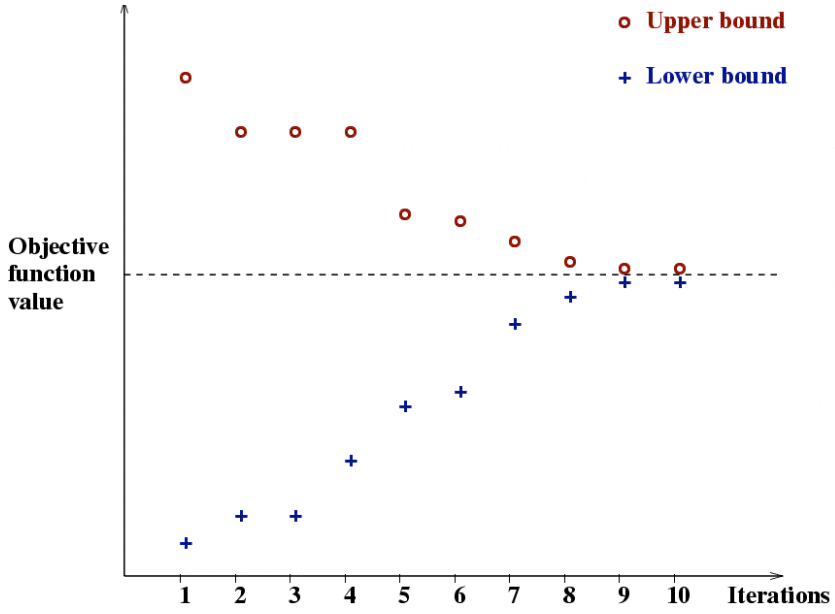
$$E_{\text{distance}} = \sum_{j \in \text{upper}} \begin{cases} A_j (d_j - d_j^{\text{U}})^2 & \text{if } d_j > d_j^{\text{U}} \\ 0 & \text{otherwise} \end{cases}$$
$$+ \sum_{j \in \text{lower}} \begin{cases} A_j (d_j - d_j^{\text{L}})^2 & \text{if } d_j < d_j^{\text{L}} \\ 0 & \text{otherwise} \end{cases}$$

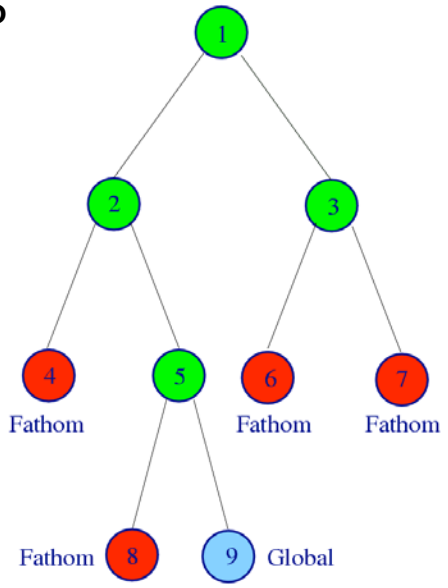
# αBB Global Optimization

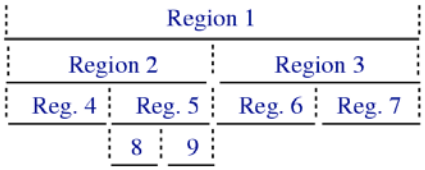
- Based on a branch-and-bound framework
- Upper bound on the global solution is obtained by solving the full nonconvex problem to local optimality
- Lower bound is determined by solving a valid convex underestimation of the original problem
- Convergence is obtained by successive subdivision of the region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

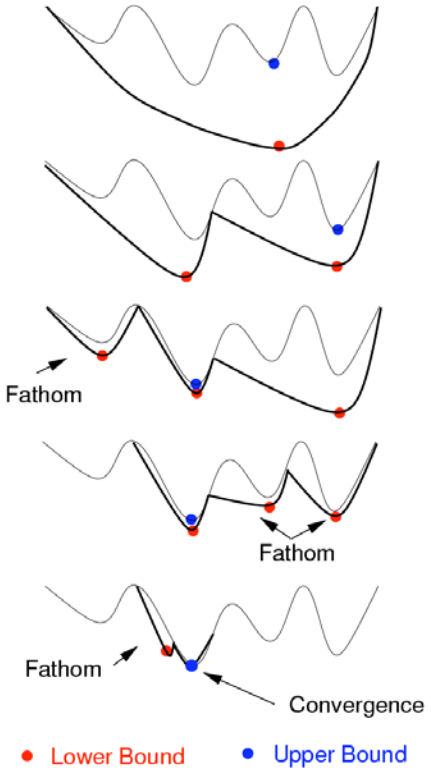
$$\begin{aligned} \min_{\mathbf{x}} \quad & f(\mathbf{x}) \\ \text{s.t.} \quad & \mathbf{h}(\mathbf{x}) = \mathbf{0} \\ & \mathbf{g}(\mathbf{x}) \leq \mathbf{0} \\ & \mathbf{x} \in \mathbf{X} \subseteq \mathcal{R}^n \end{aligned}$$

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



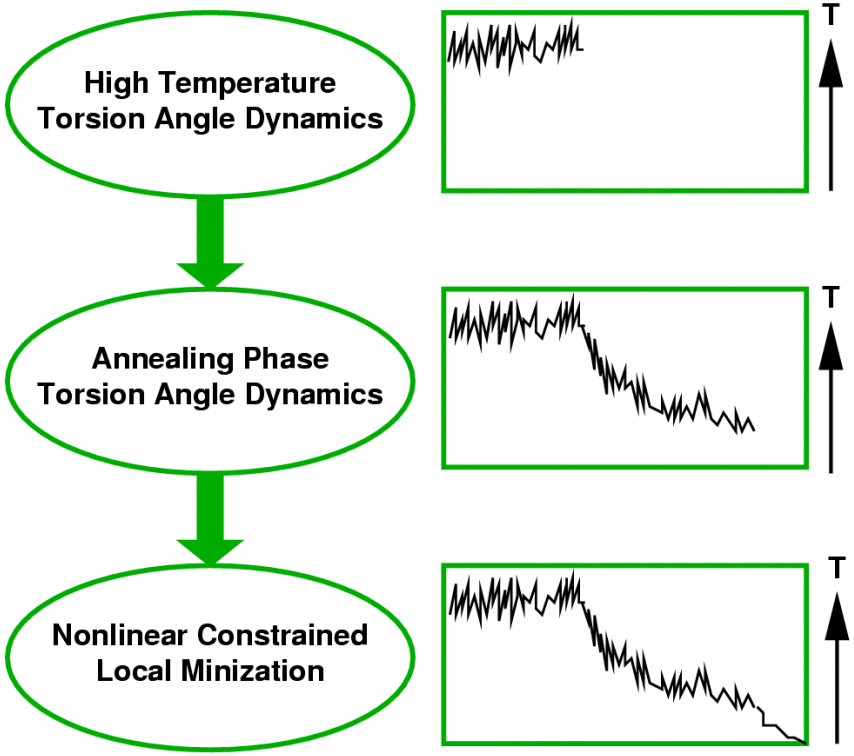




Floudas, CA and co-workers, 1995-2006

Adjiman, CS, et al. Computers and Chemical Engineering. (1998a,b)

# Torsion Angle Dynamics

- Why? Difficult to identify low energy feasible points
- Fast evaluation of steric based force field
- Unconstrained formulation with penalty functions

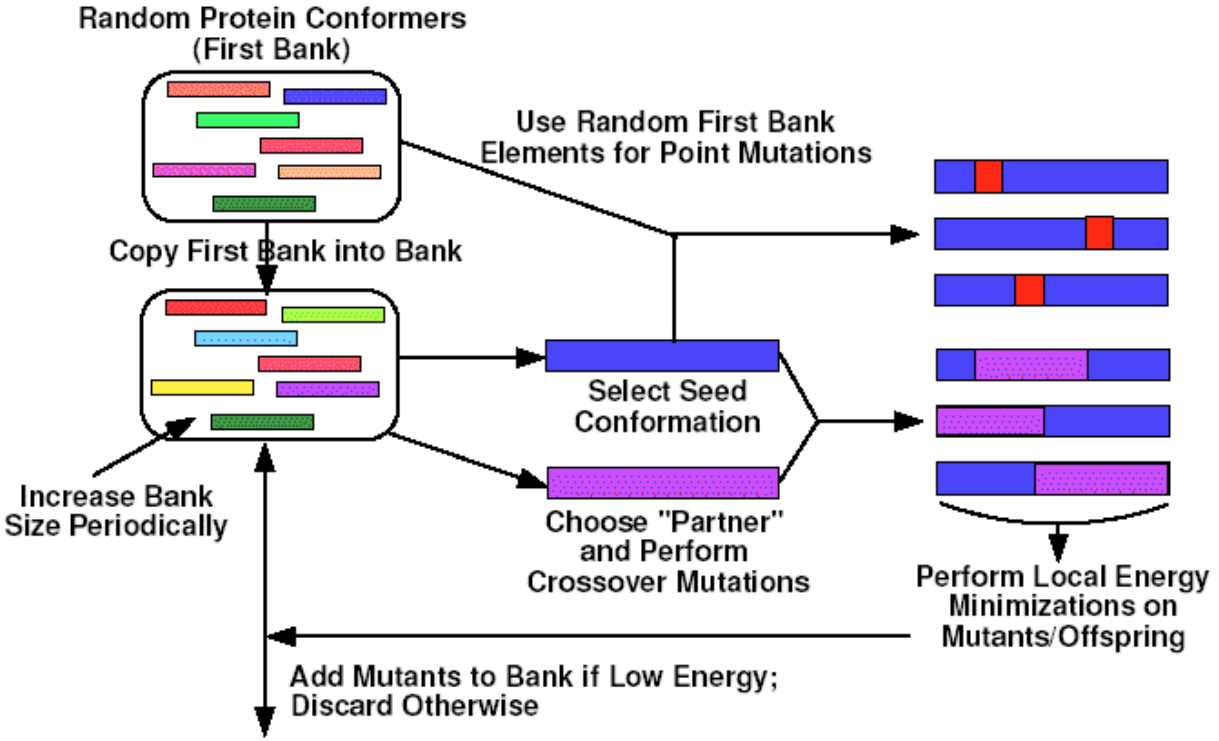


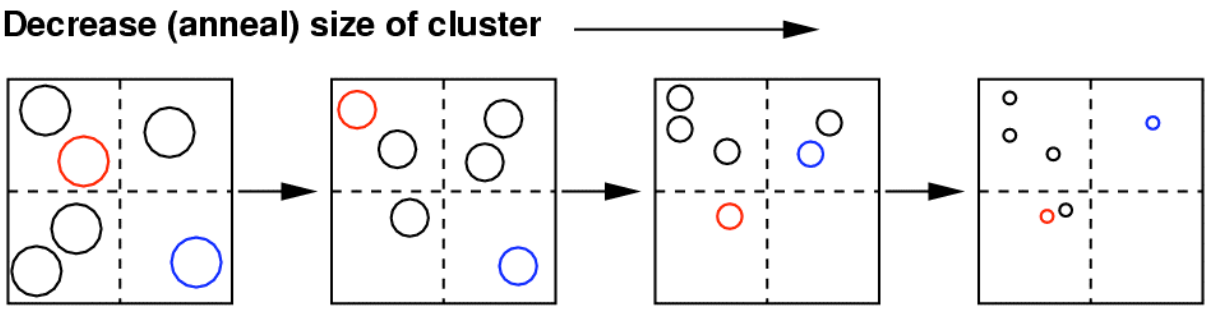

- Implemented by solving equations of motion as preprocessing for each constrained minimization

Guntert, P, et al. Journal of Molecular Biology. (1997)
Klepeis, JL, et al. Journal of Computational Chemistry. (1999)
Klepeis, JL and Floudas, CA. Computers and Chemical Engineering. (2000)

# Conformational Space Annealing

- Induce variations
  - Mutations
  - Crossovers
- Subject to local energy minimization
- Anneal through the gradual reduction of space




Scheraga and co-workers, 1997-2006.

Lee, JH, et al. Journal of Computational Chemistry. (1997)

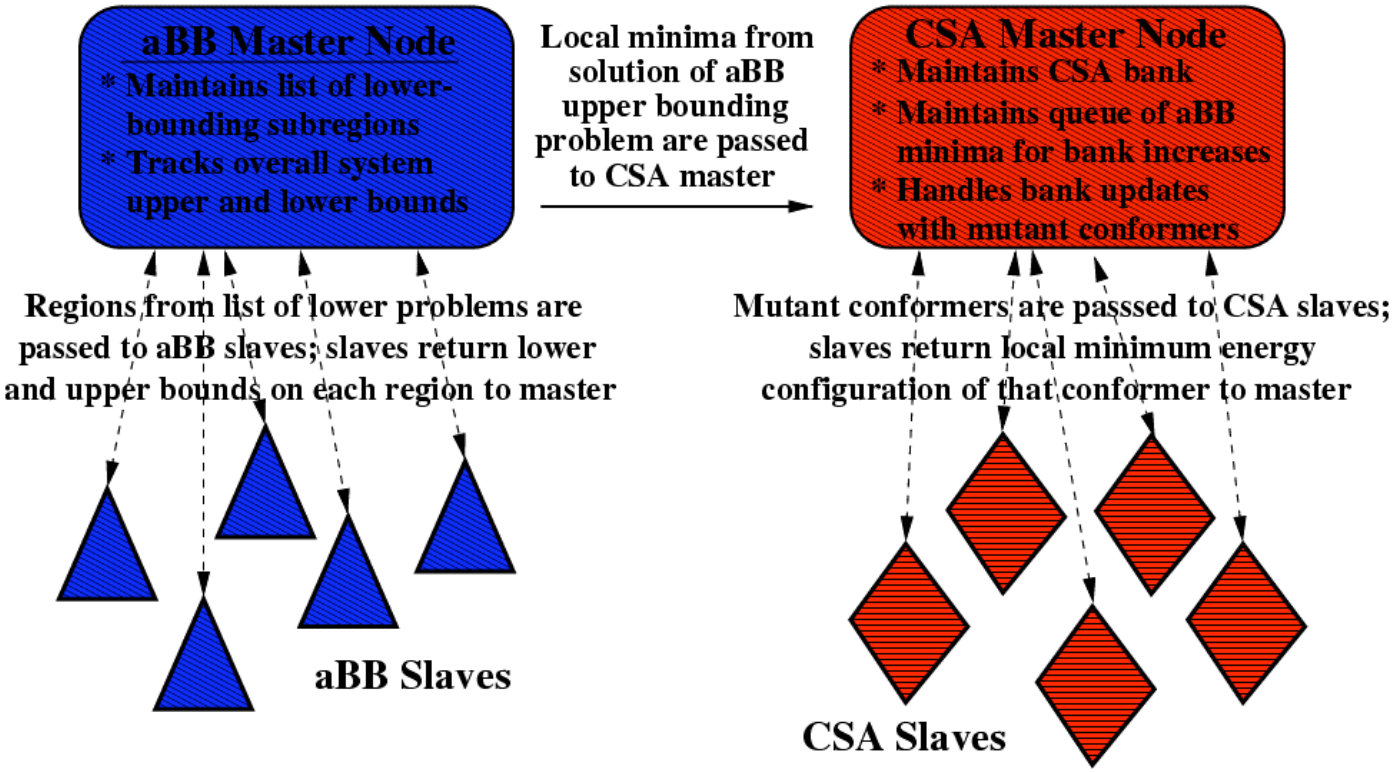
# Tertiary Structure Prediction

- Hybrid global optimization approach
  - αBB deterministic global optimization
  - Conformational Space Annealing (CSA)
- Modifications
  - Streamlined parallel implementation
  - Integrated a rotamer optimization stage for quick energetic improvements
  - Improved initial point selection using a torsion angle dynamics based annealing procedure from CYANA*

*Guntert, P, et al. Journal of Molecular Biology. (1997)

# Global Optimization: Alternating Hybrid

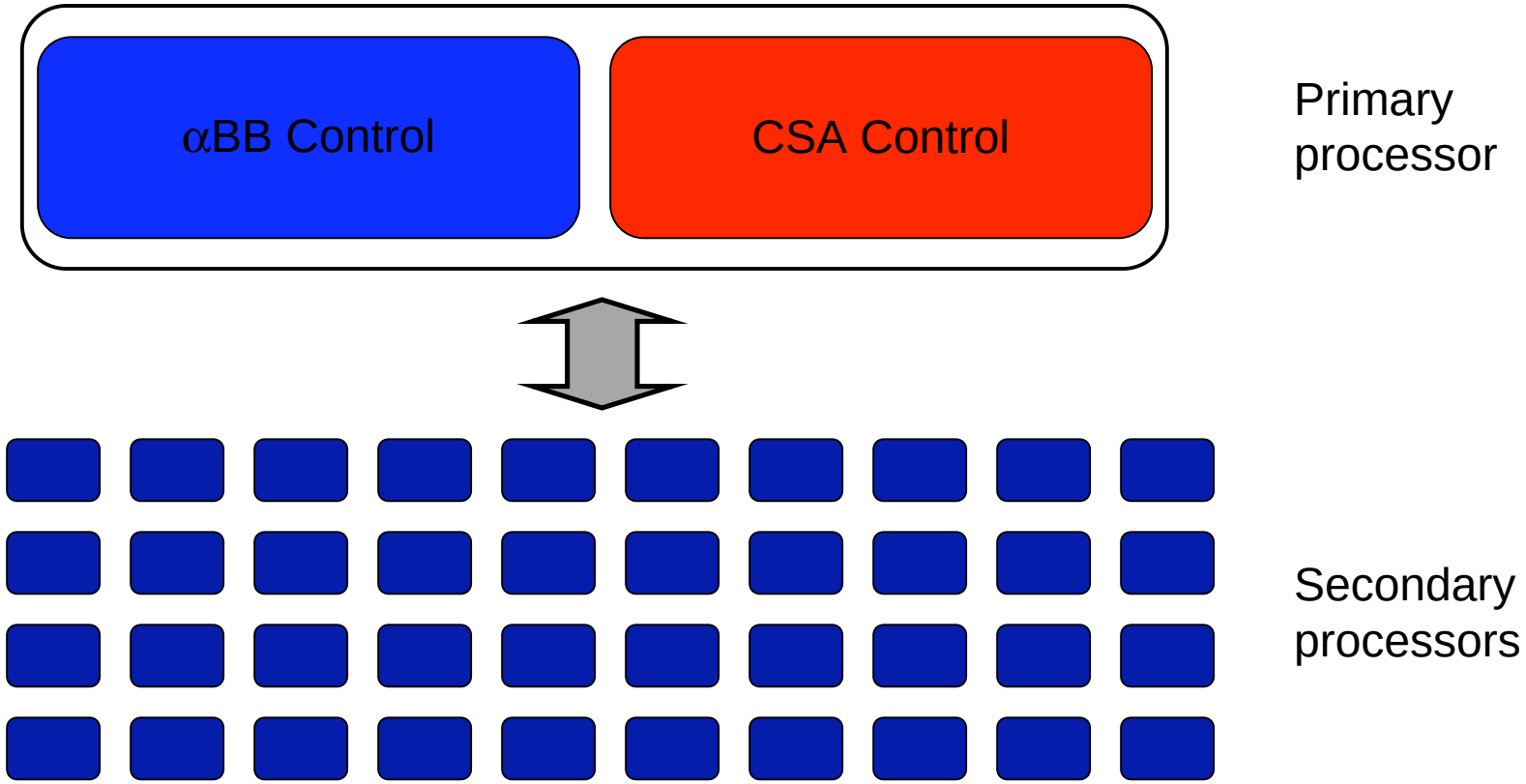
- The αBB and CSA algorithms have complementary strengths and drawbacks
- Implement hybrid algorithm to capture strengths of both
- Parallelize by dividing problem, assigning subproblems




Klepeis, JL, et al. Biophysical Journal. (2003)

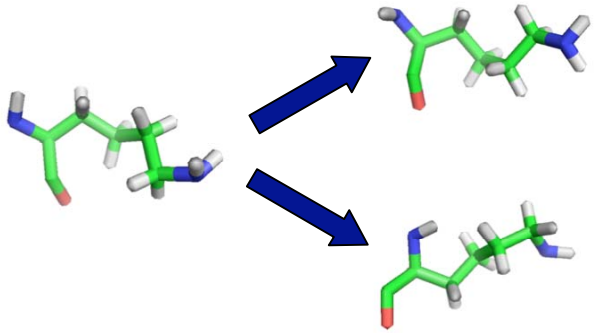
# Alternating Hybrid: Implementation

- All secondary nodes begin performing αBB iterations
- Once the CSA bank is full, CSA takes control of a subset of secondary nodes

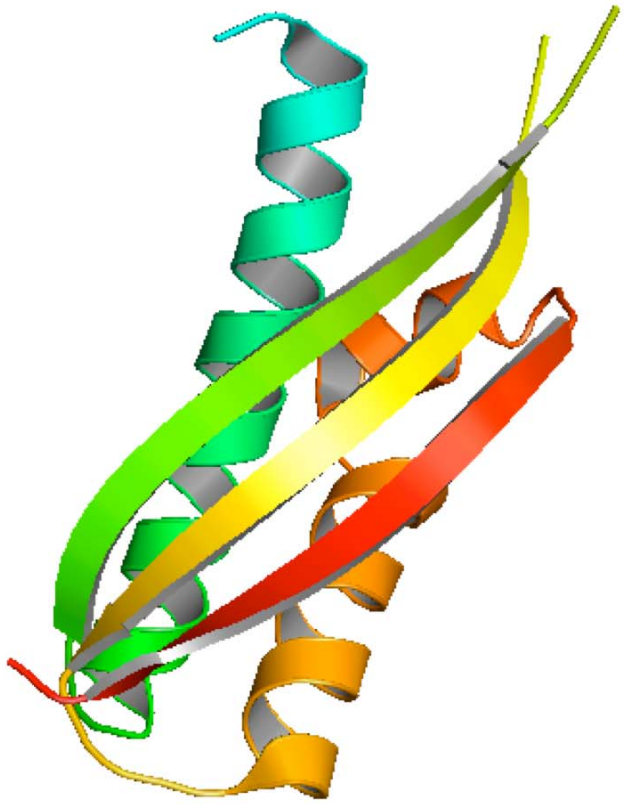


# Rotamer Side Chain Optimization

- Side chain packing is crucial to the stability and specificity of the native state
- Rotamer optimization is a quick way to alleviate steric clashes
- Better starting point for constrained nonlinear minimization



# Tertiary structure prediction

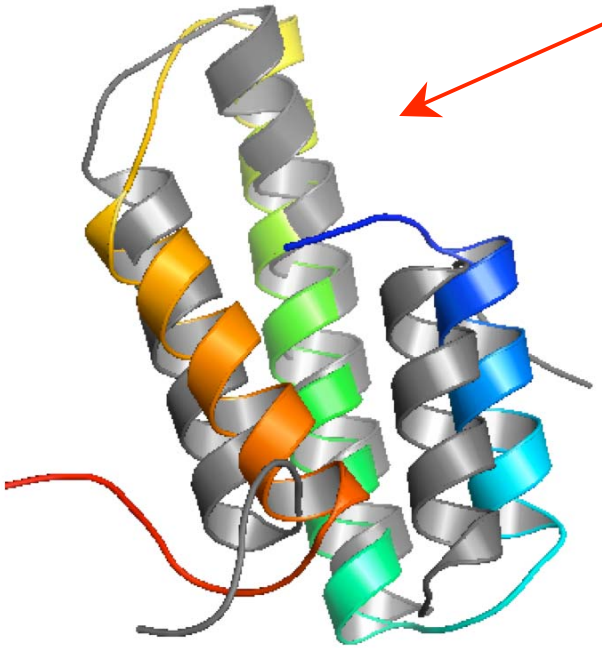


Results

# Results – Tertiary Structure Prediction

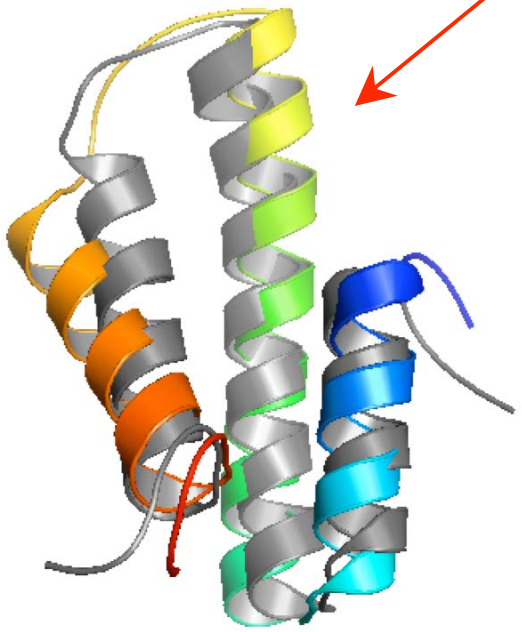
- PDB: 1nre

Energy -1395.48
RMSD 6.63



Lowest energy predicted structure of 1nre (color) versus native 1nre (gray)

Energy -1340.45
RMSD 3.52

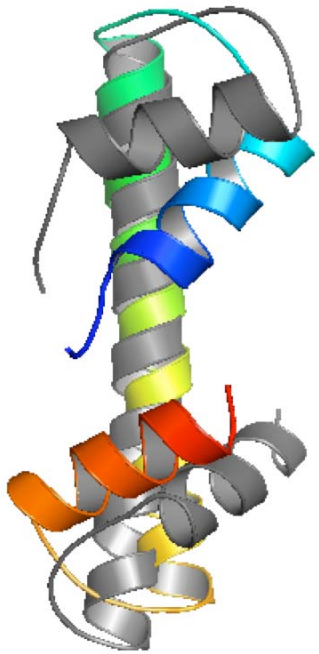


Lowest RMSD predicted structure of 1nre (color) versus native 1nre (gray)

# Results – Tertiary Structure Prediction

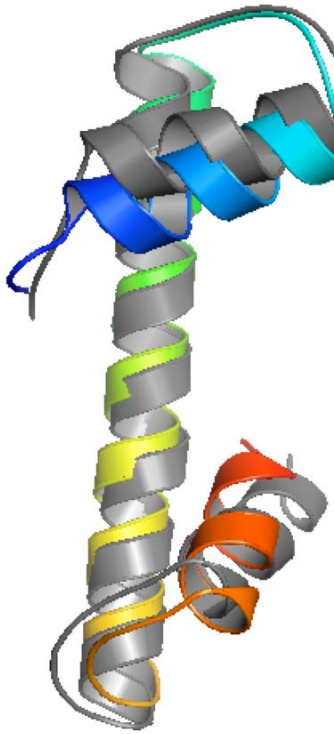
- PDB: 1hta

Energy -941.02
RMSD 6.70



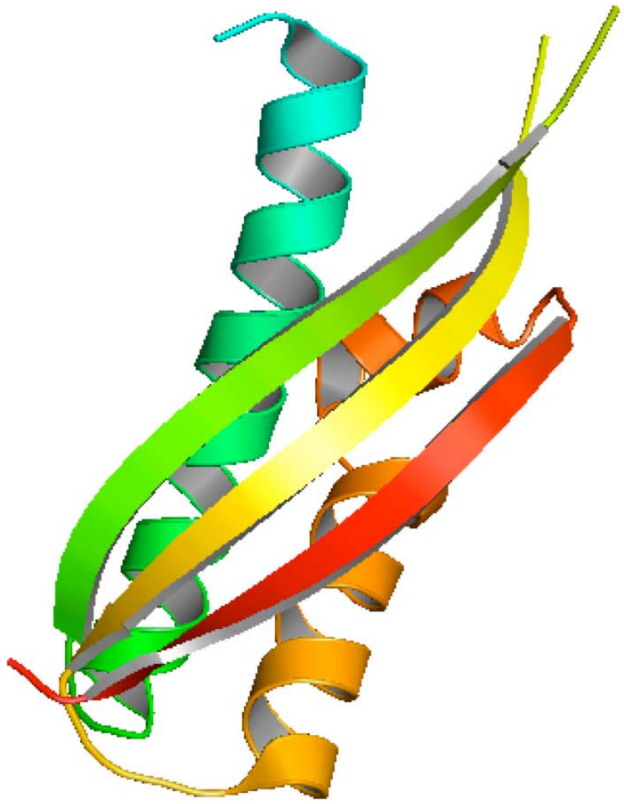
Lowest energy predicted structure of 1hta (color) versus native 1hta (gray)

Energy -915.57
RMSD 2.58



Lowest RMSD predicted structure of 1hta (color) versus native 1hta (gray)

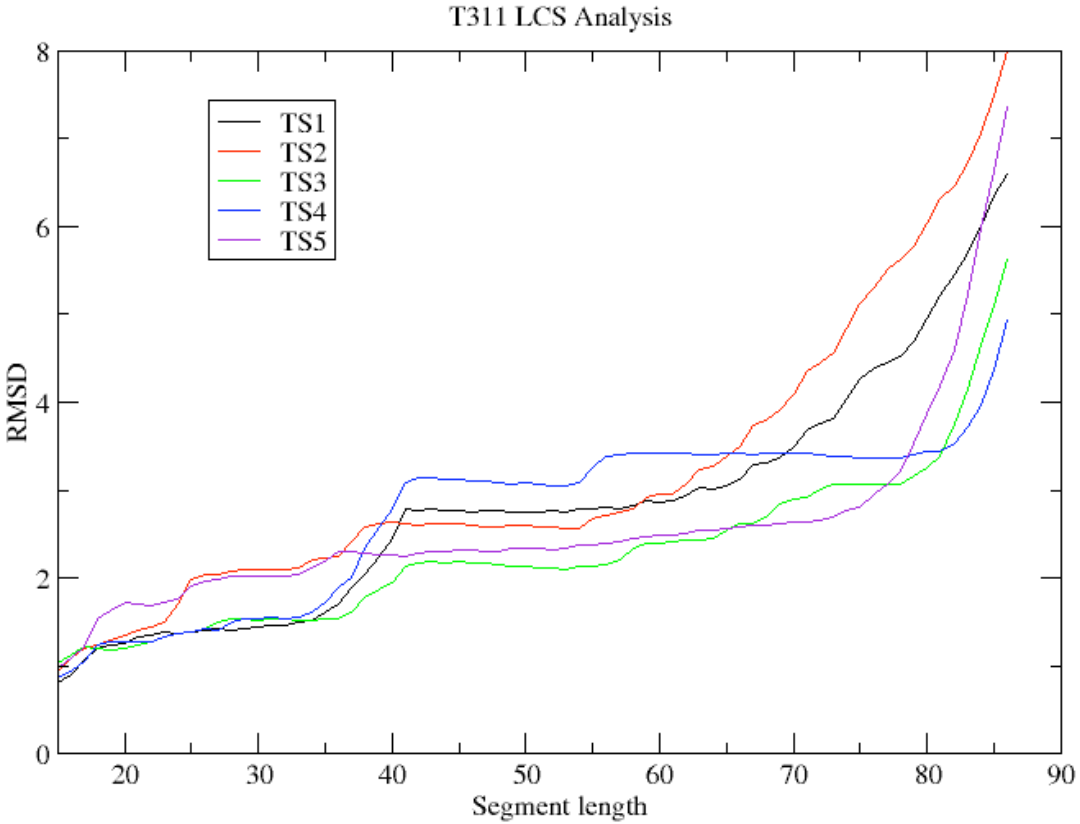
# Tertiary structure prediction



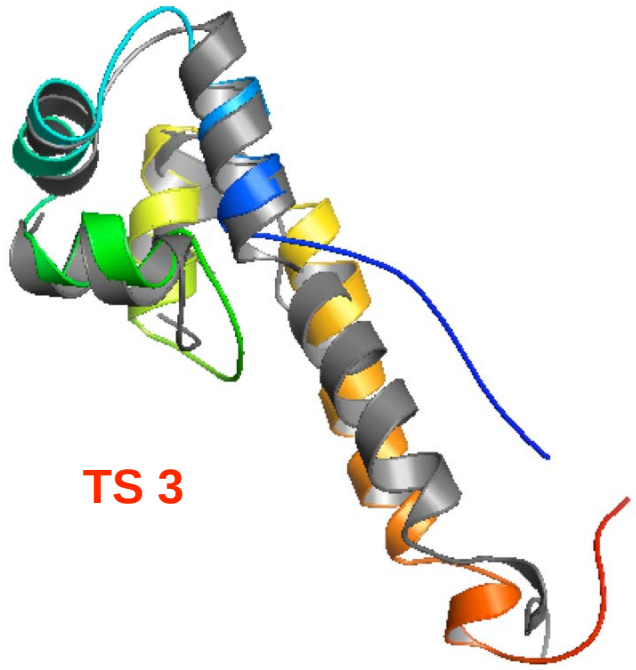
Blind studies

# CASP7 Results – T311 (87/97 aa)

- PSIPRED used for $\alpha$-helical prediction
- A small number of loose, homology based



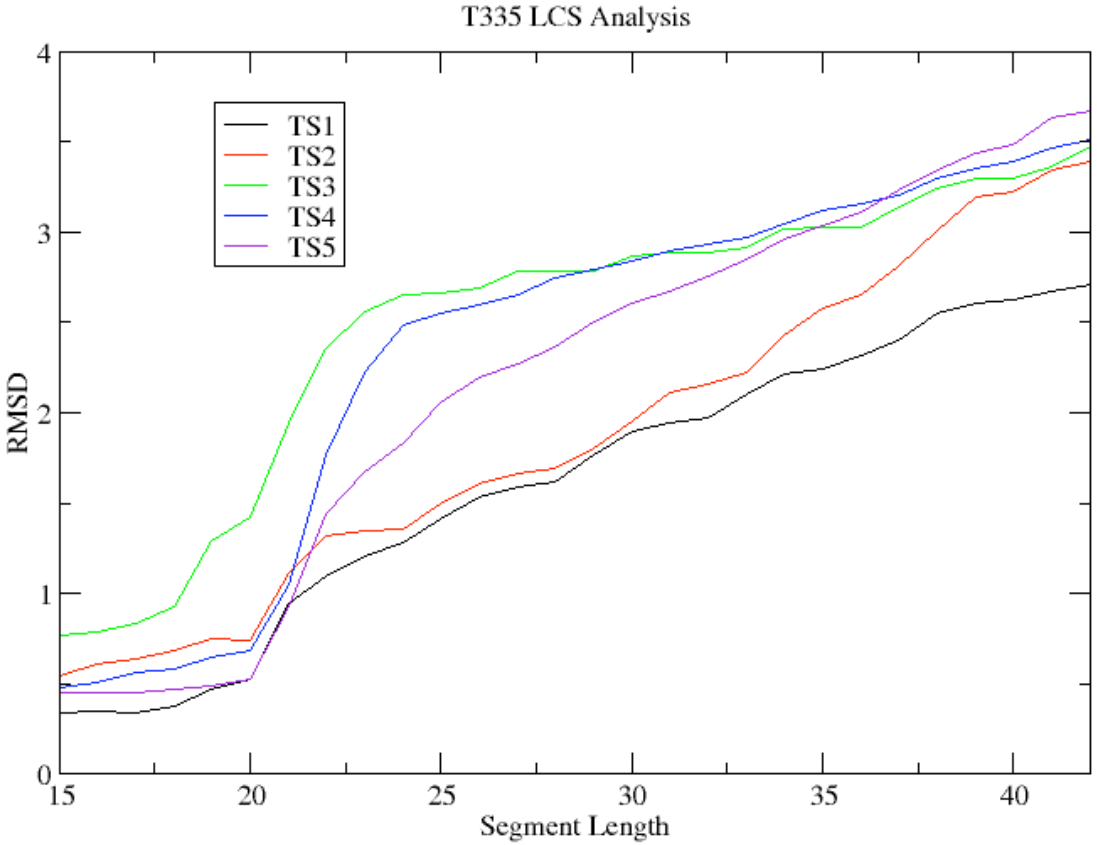

RMSD 5.75

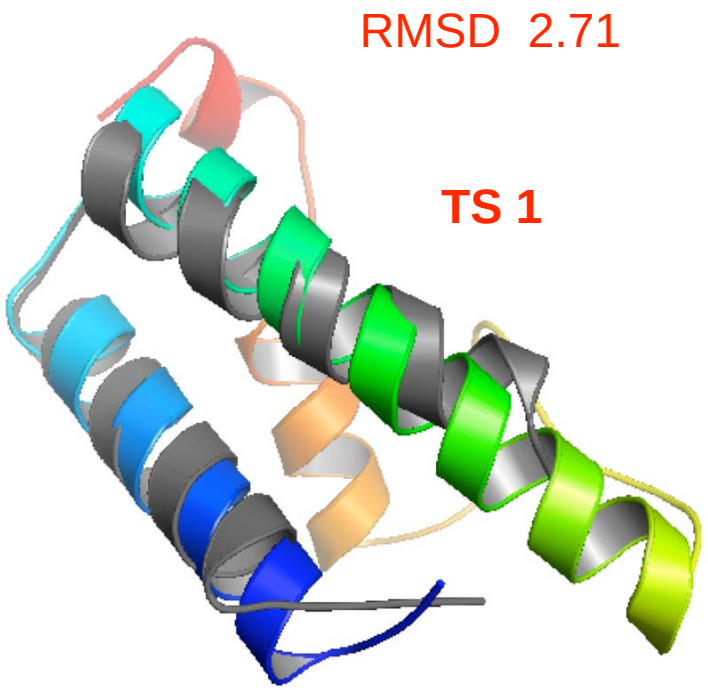


TS 3

# CASP7 Results – T335 (42/85 aa)

- PSIPRED used for α-helical prediction
- Distance constraints imposed based on α-helical

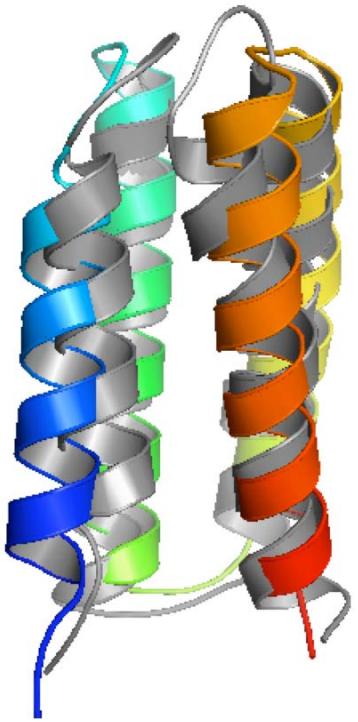


# Results – Blind Tertiary Structure Prediction

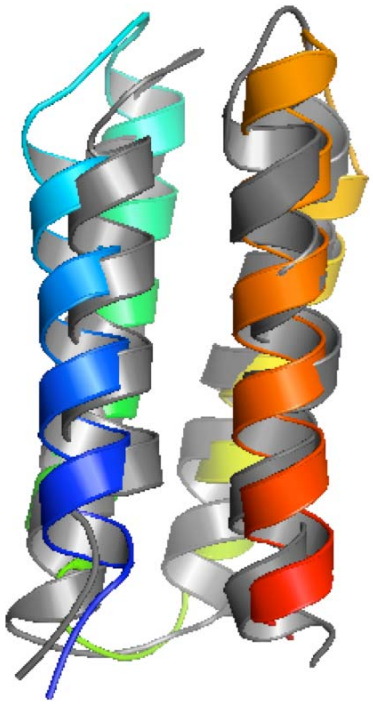
- S836

Energy -1740.11
RMSD 2.84



Lowest energy predicted structure of s836 (color) versus native s836 (gray)

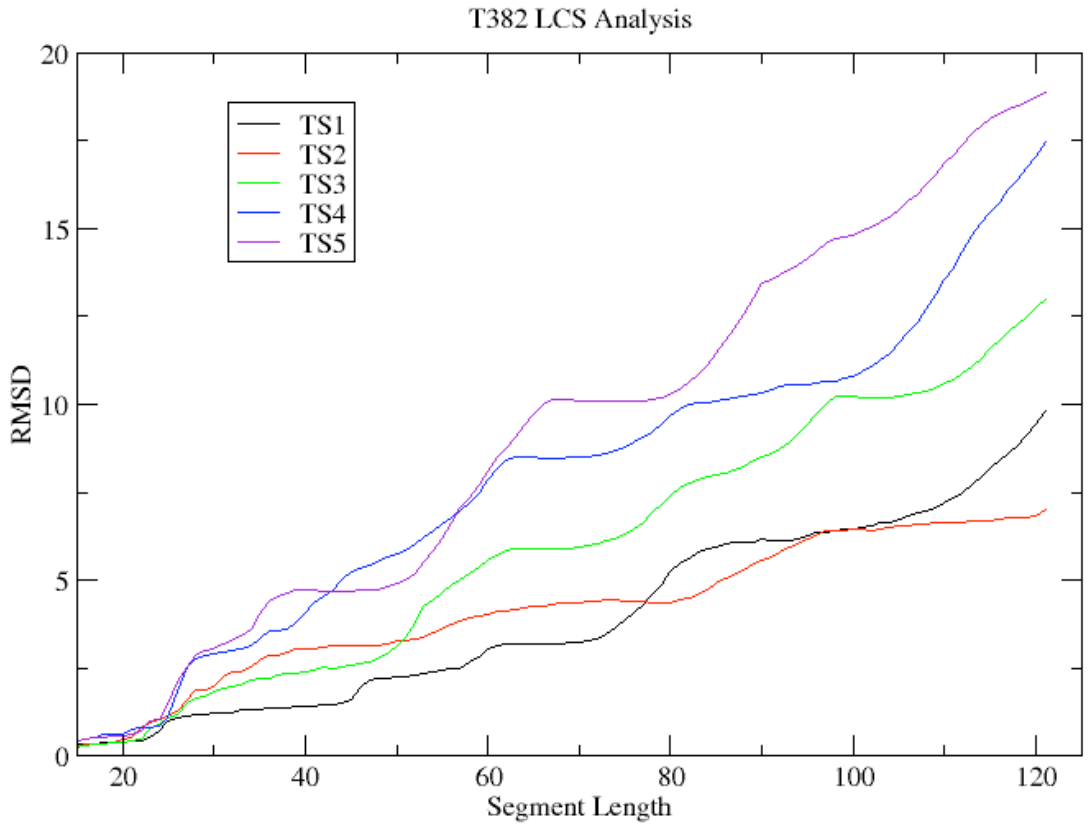
Energy –1697.88
RMSD 2.39

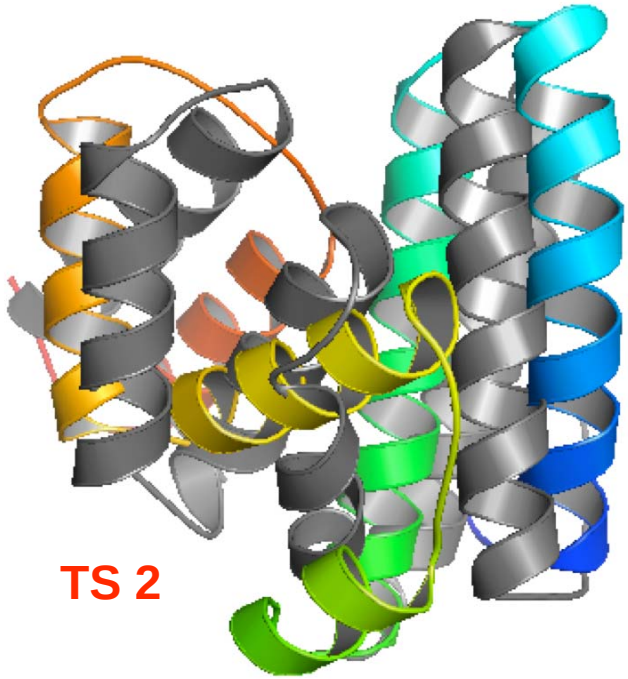


Lowest RMSD predicted structure of s836 (color) versus native s836 (gray)

# CASP7 Results ,T382 (121/123 aa)

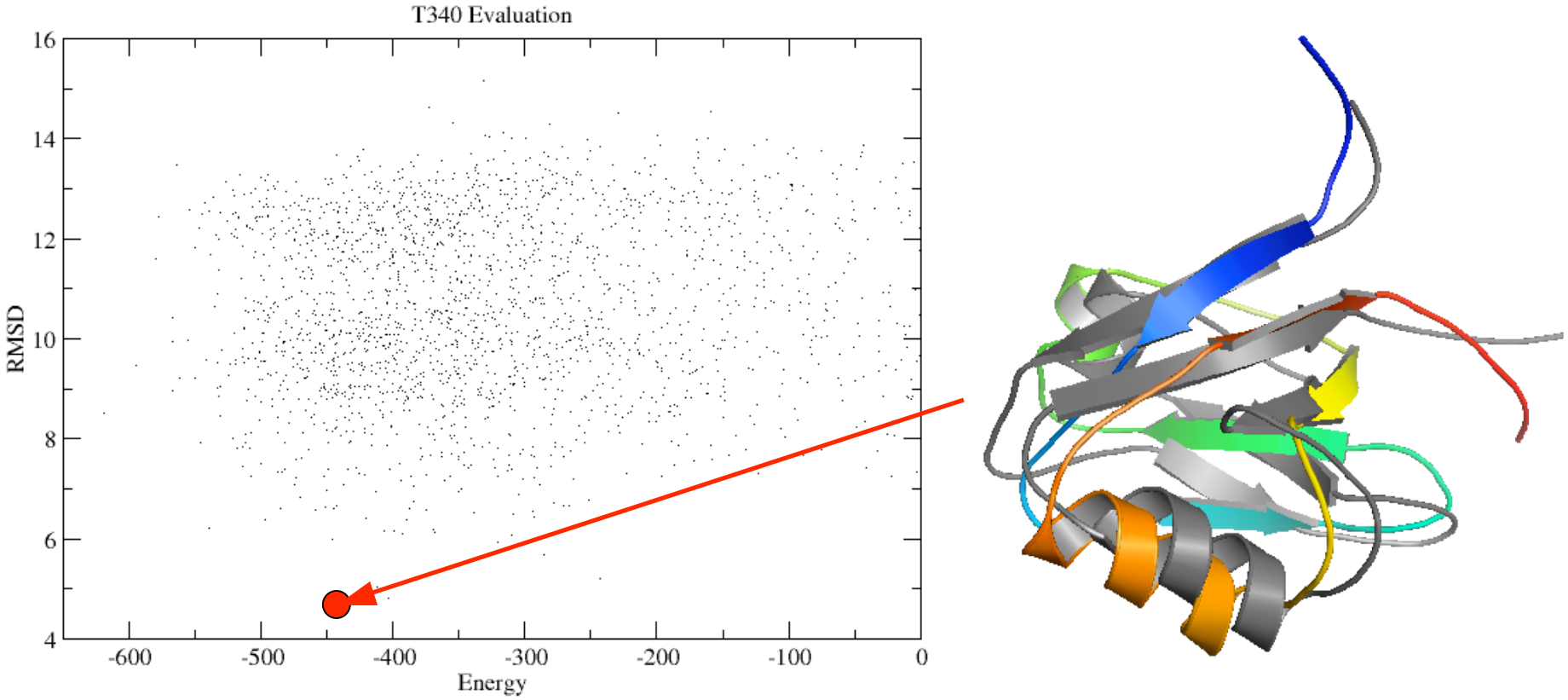
- PSIPRED, PROFsec used for α-helical prediction
- Distance constraints imposed based on α-helical topology prediction results



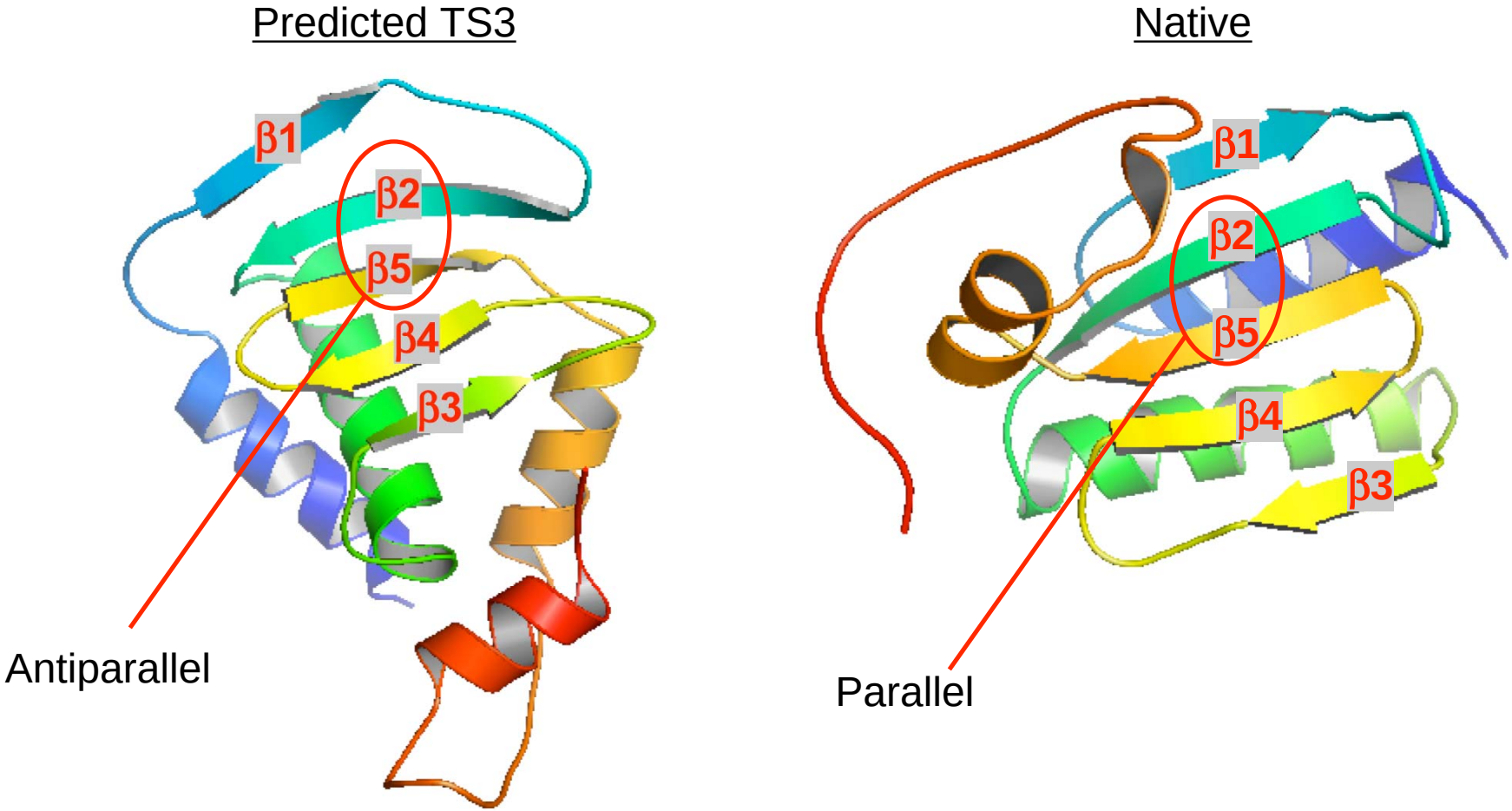
# CASP7 Results – T340 (90 aa)

- Selection by energy alone may not identify the lowest RMSD structure



# CASP7 Results-T354 (120/130 aa)

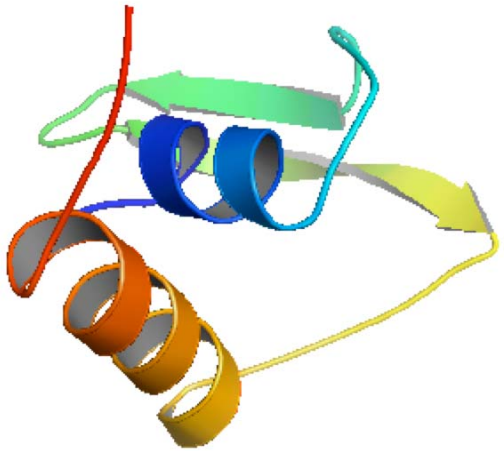
- Incorrect topology predictions can misdirect global conformational search



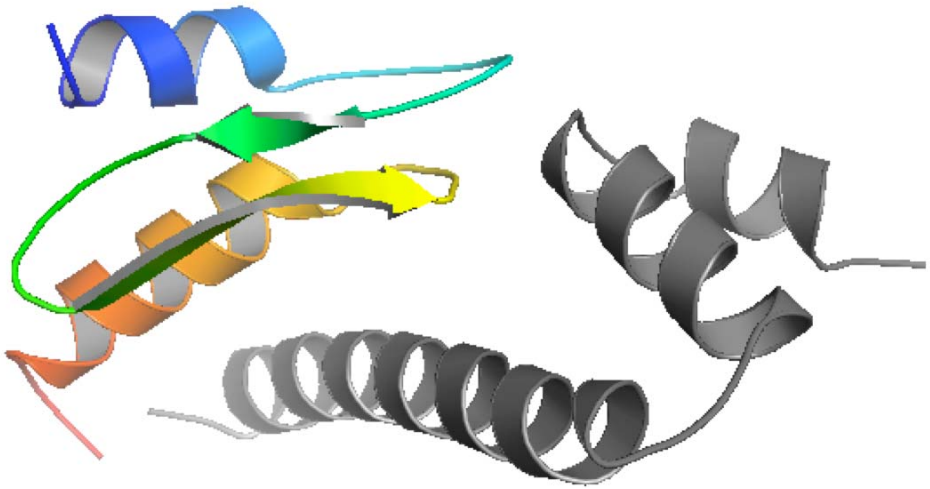
# CASP7 Results – T351 (60/117 aa)

- Overall RMSD can be deceiving, hiding a correct topology prediction

Native



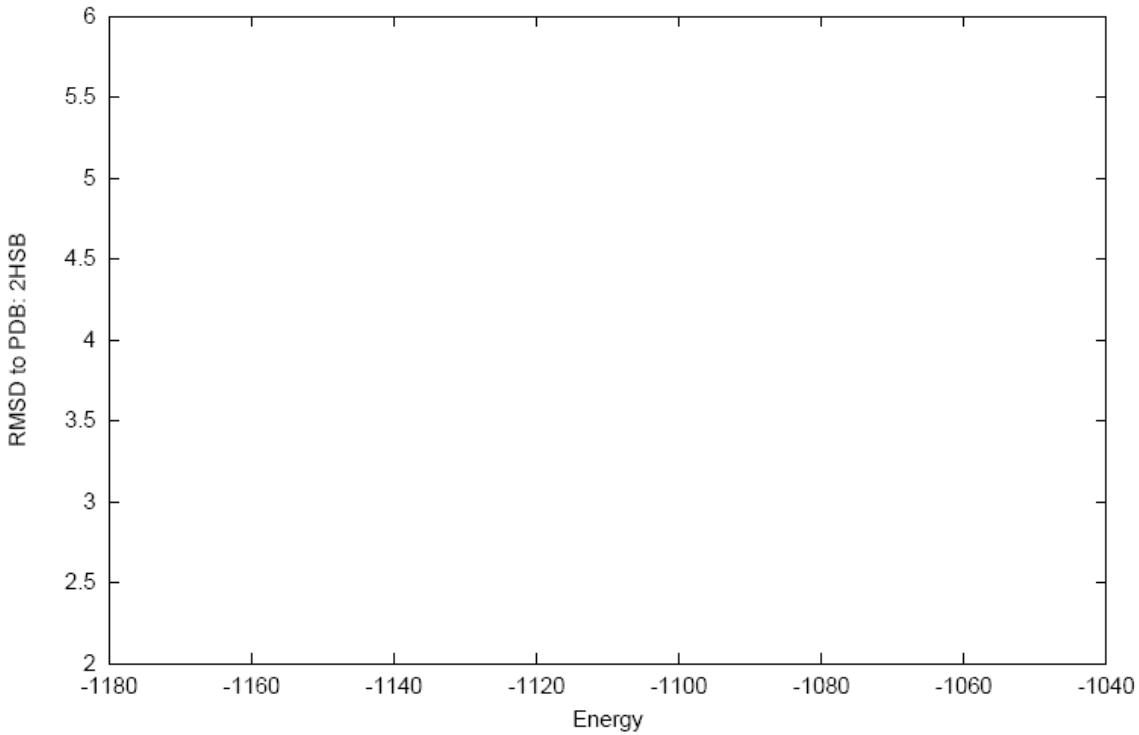
Predicted TS1

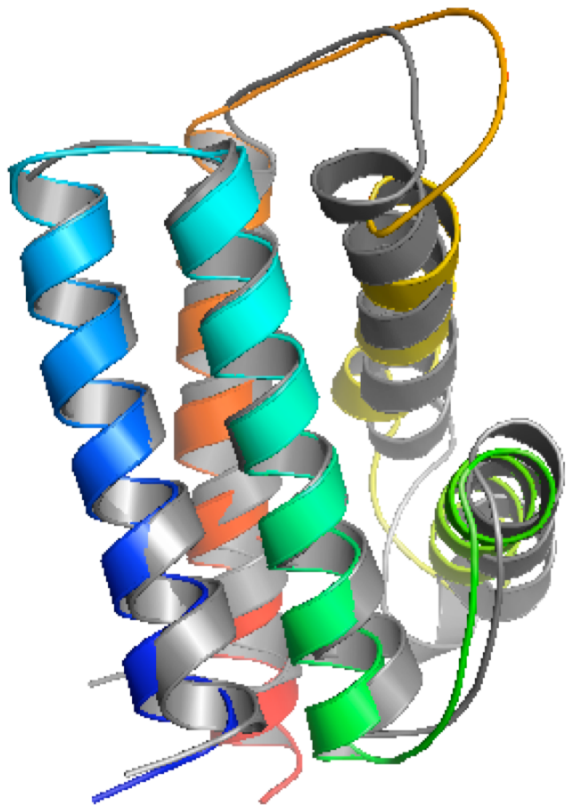


9.69 RMSD

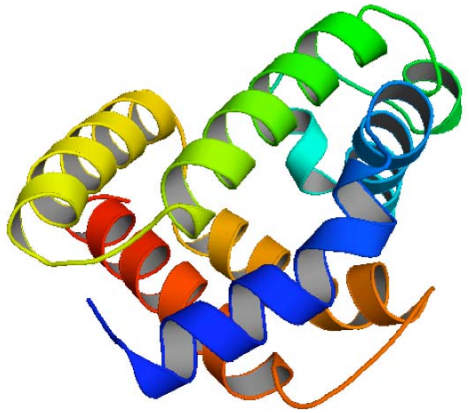
# CASP7 – T367 Comparison

- PSIPRED used for α-helical prediction
- Tight distance constraints imposed based on α-helical topology prediction results



# Alpha-helical Topology and Tertiary Structure Prediction of Globular Proteins

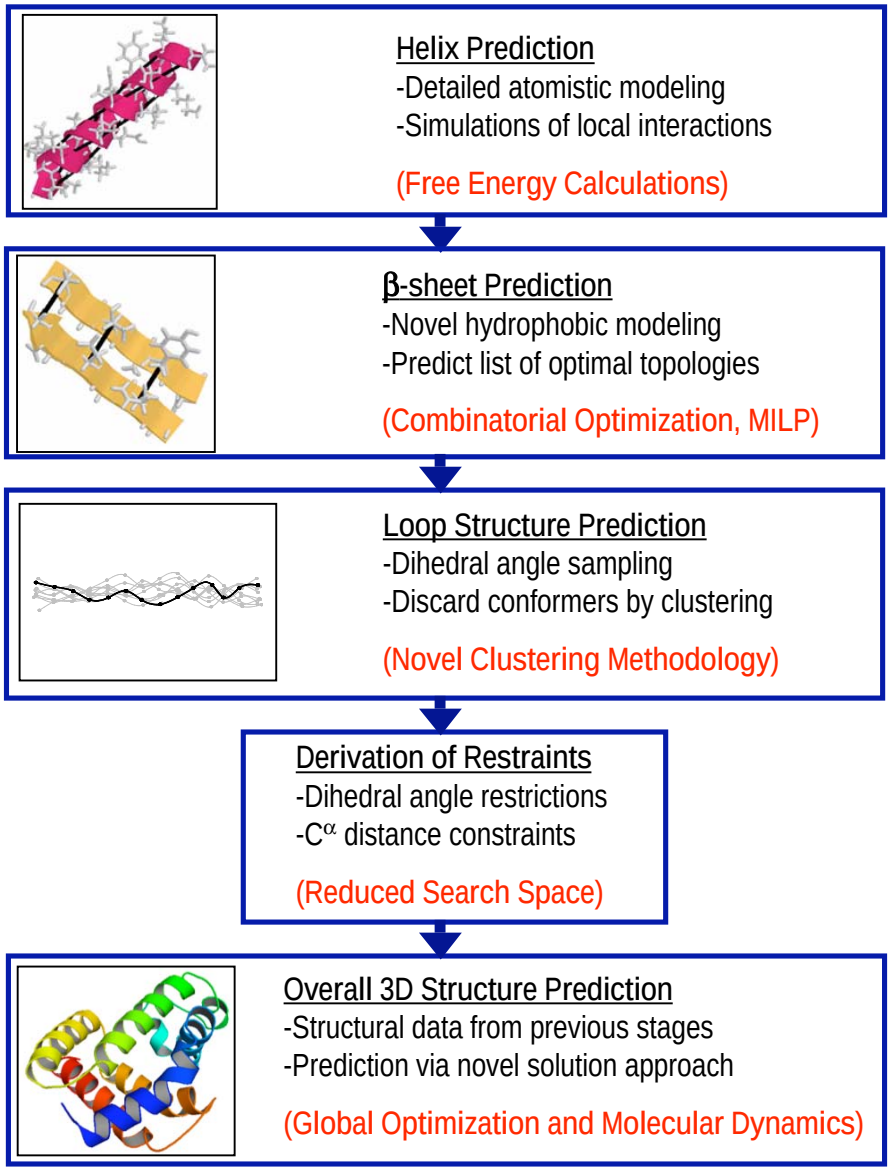


**McAllister S.R., Mickus B.E., J.L. Klepeis, and C.A. Floudas, "A Novel Approach for Alpha-Helical Topology Prediction in Globular Proteins: Generation of Interhelical Restraints", Proteins, 65, 930-952 (2006).**

# Outline

- Predicting α-helical contacts
  - Probability development
  - Model
  - Results
- Predicting α-helical contacts in α/β proteins
  - Distance bounding
  - Model
  - Results
- Structure prediction of α-helical proteins
  - Framework
  - Results

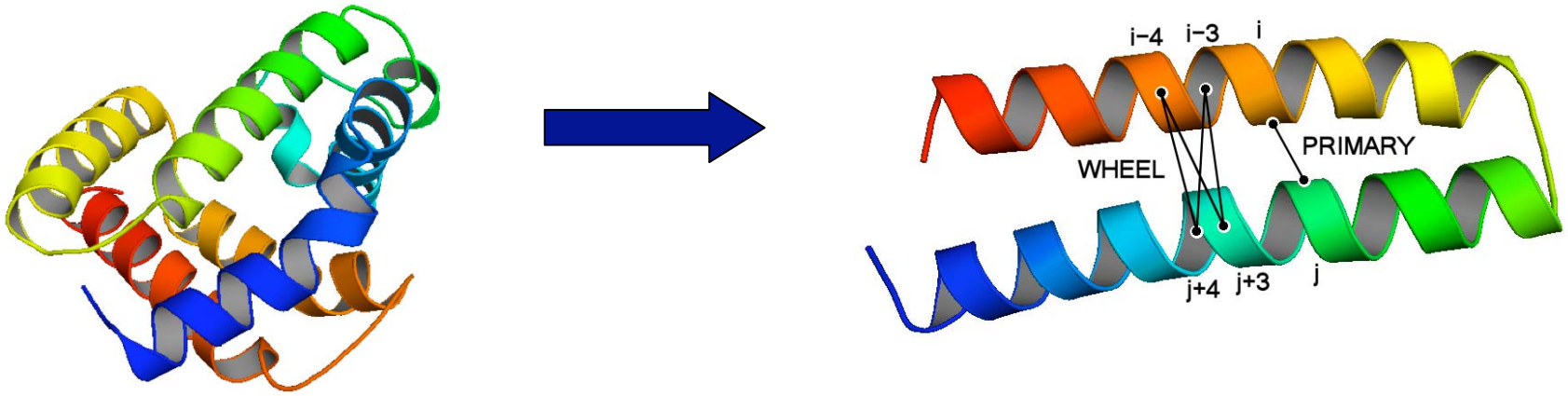
# ASTRO-FOLD




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

# Overview

- **Problem**
  - Topology prediction of globular $\alpha$-helical proteins
- **Approach**
- **Thesis: Topology is based on certain Inter-helical Hydrophobic to Hydrophobic Contacts**
  - Create a dataset of helical proteins
  - Develop inter-helical contact probabilities
  - Apply two novel mixed-integer optimization models (MILP)
    - Level 1 - PRIMARY contacts
    - Level 2 - WHEEL contacts




McAllister, Mickus, Klepeis, Floudas. Proteins. 2006, 65:930-952.

# Dataset Selection

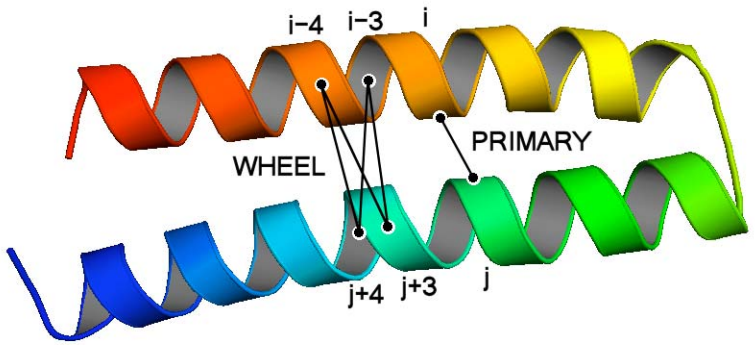
- **Protein Sources**
  - –229 PDBSelect25[1] database
  - –62 CATH[2] database
  - –20 Zhang et al.[3]
  - –7 Huang et al.[4]
- **Restrictions**
  - –No β-sheets, at least 2 α-helices
  - –No highly similar sequences
- **Dataset**
  - –318 proteins in the database set

[1]Hobohm, U. and C.Sander. *Prot Sci* **3** (1994) 522
[2]Orengo, C.A. et al. *Structure* 5 (1997) 1093.
[3]Zhang, C. et al. *PNAS* 99 (2002) 3581.
[4]Huang, E.S. et al. *J Mol Biol* 290 (1999) 267.

McAllister, Mickus, Klepeis, Floudas. Proteins. 2006, 65:930-952.

# Probability Development

- **Contact Types**
  - **PRIMARY** contact
    - Minimum distance **hydrophobic** **contact** between 4.0 Å and 10.0 Å
  - **WHEEL** contact
    - Only WHEEL position **hydrophobic** **contacts** between 4.0 Å and 12.0 Å
- Classified as **parallel** or **antiparallel** contacts




McAllister, Mickus, Klepeis, Floudas. Proteins. 2006, 65:930-952.

# Model Overview

- **Formulation: Maximize inter-helical residue-residue contact probabilities**
  - **Binary variables** indicate **antiparallel** helical contact
  - **Binary variables** indicates residue contact
- **Goal:** Produce a **rank-ordered list** of the most likely helical contacts
  - Contacts used to **restrict conformational space** explored during protein tertiary structure prediction

McAllister, Mickus, Klepeis, Floudas. Proteins. 2006, 65:930-952.

# Pairwise Model Objective

- **Level 1 Objective**
  - **–Maximize probability of pairwise residue-residue contacts**

$$\max \quad \sum_m \sum_n y^a_{mn} \cdot \sum_i \sum_j w^{mn}_{ij} \cdot p^a_{ij;mn}$$

$$+ \quad \sum_m \sum_n y^p_{mn} \cdot \sum_i \sum_j w^{mn}_{ij} \cdot p^p_{ij;mn}$$

$$y^a_{mn}, y^p_{mn}, w^{mn}_{ij} = \{0, 1\}$$

McAllister, Mickus, Klepeis, Floudas. Proteins. 2006, 65:930-952.

# Pairwise Model Constraints

- **Level 1 Constraints**
  - **At most one contact per position**

$$\sum_{j;j>i} w_{ij} + \sum_{j;j<i} w_{ij} \leq 1$$

  - **Helix-helix interaction direction**

$$y^a_{mn} + y^p_{mn} \leq 1 \quad \forall(m,n)$$

  - **Linking i**

$$w^{mn}_{ij} \leq y^a_{mn} + y^p_{mn}$$

$$y^a_{mn} + y^p_{mn} - \sum_i \sum_j w^{mn}_{ij} \leq 0$$

McAllister, Mickus, Klepeis, Floudas. Proteins. 2006, 65:930-952.

# Pairwise Model Constraints

- **Level 1 Constraints**
  - **–Restrict number of contacts between a given helix pair (MAX_CONTACT)**

$$\sum_i \sum_j w_{ij}^{mn} \leq max\_contact \cdot (y_{mn}^a + y_{mn}^p) \quad \forall (m,n)$$

  - **–Vary the number of helix-helix interactions (SUBTRACT)**

$$\sum_m \sum_n (y_{mn}^a + y_{mn}^p) \leq \left( \sum_m counth(m)/2 \right) - subtract$$

McAllister, Mickus, Klepeis, Floudas. Proteins. 2006, 65:930-952.

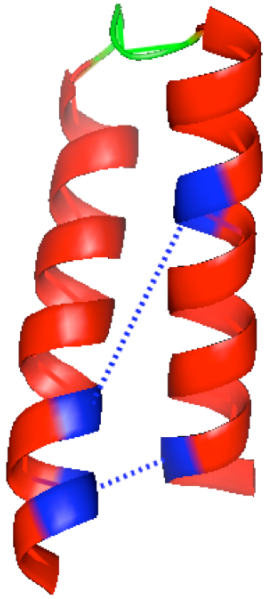
# Pairwise Model Constraints

- **Level 1 Constraints**
  - **Allow for and Limit helical kinks**

$$w_{ij}^{mn} + w_{i'j'}^{mn} \leq 1$$

$$\forall(i, i', j, j') : \quad |\text{diff}(i, i')| - |\text{diff}(j, j')| \leq 2$$

$$\text{or either } |\text{diff}(i, i')| \leq 5 \text{ or } |\text{diff}(j, j')| \leq 5$$



McAllister, Mickus, Klepeis, Floudas. Proteins. 2006, 65:930-952.

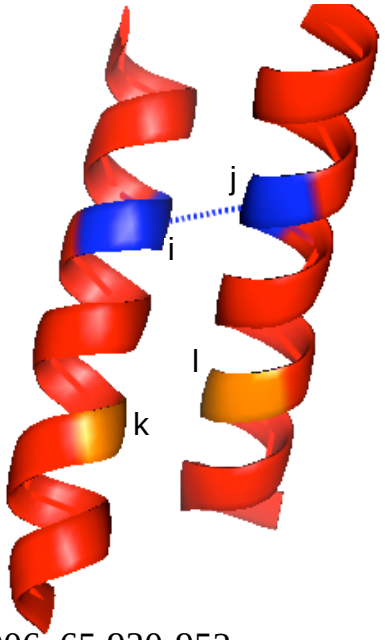
# Pairwise Model Constraints

- **Level 1 Constraints**
  - **–Consistent numbering**

$$w_{ij}^{mn} + w_{i'j'}^{mn} + y_{mn}^{a} \leq 2 \quad \forall (i,j,i',j') : \quad i' > i, j' > j \text{ and}$$

$$|i' - i| < |j' - j| + 3 \text{ or}$$

$$|i' - i| > |j' - j| - 3$$




McAllister, Mickus, Klepeis, Floudas. Proteins. 2006, 65:930-952.

# Pairwise Model Constraints

- **Feasible topologies**

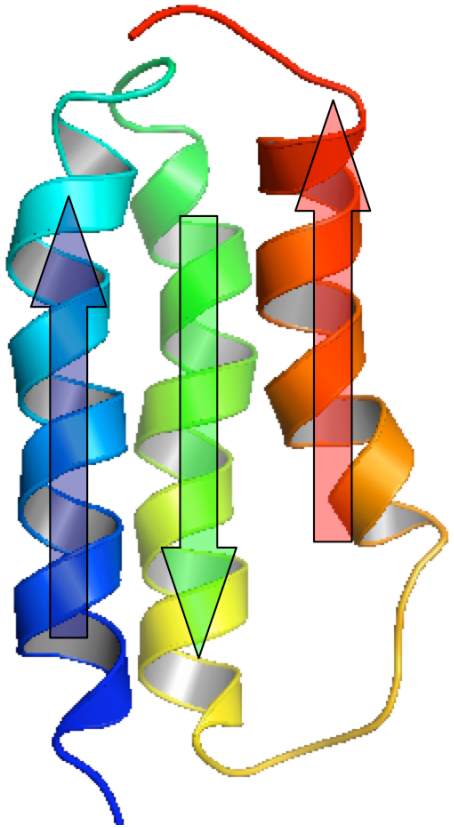
$$y^p_{mn} + y^a_{np} + y^p_{mp} \leq 2 \quad \forall (m,n,p) : m \neq n \neq p, nhel \geq 3$$

$$y^a_{mn} + y^p_{np} + y^p_{mp} \leq 2 \quad \forall (m,n,p) : m \neq n \neq p, nhel \geq 3$$

$$y^p_{mn} + y^p_{np} + y^p_{mp} \leq 2 \quad \forall (m,n,p) : m \neq n \neq p, nhel \geq 3$$

$$y^a_{mn} + y^a_{np} + y^a_{mp} \leq 2 \quad \forall (m,n,p) : m \neq n \neq p, nhel \geq 3$$

**1** **1**



# Pairwise Model Objective

- **Level 2 Objective**
  - **–Maximize the sum of predicted wheel probabilities**

$$\max \quad \sum_{m,n}\sum_{i,j}\sum_{k,l} w^{mn}_{kl;ij} \cdot \left[p^{a}_{kl;ij;mn} + p^{a}_{ij;kl;mn}\right] \cdot y^{a}_{mn} \cdot w^{mn}_{ij}$$

$$+ \quad \sum_{m,n}\sum_{i,j}\sum_{k,l} w^{mn}_{kl;ij} \cdot \left[p^{p}_{kl;ij;mn} + p^{p}_{ij;kl;mn}\right] \cdot y^{p}_{mn} \cdot w^{mn}_{ij}$$

$$y^{a}_{mn}, y^{p}_{mn}, w^{mn}_{ij}, w^{mn}_{kl;ij}, y^{a}_{ijmn}, y^{p}_{ijmn} = \{0, 1\}$$

McAllister, Mickus, Klepeis, Floudas. Proteins. 2006, 65:930-952.

# Pairwise Model Constraints

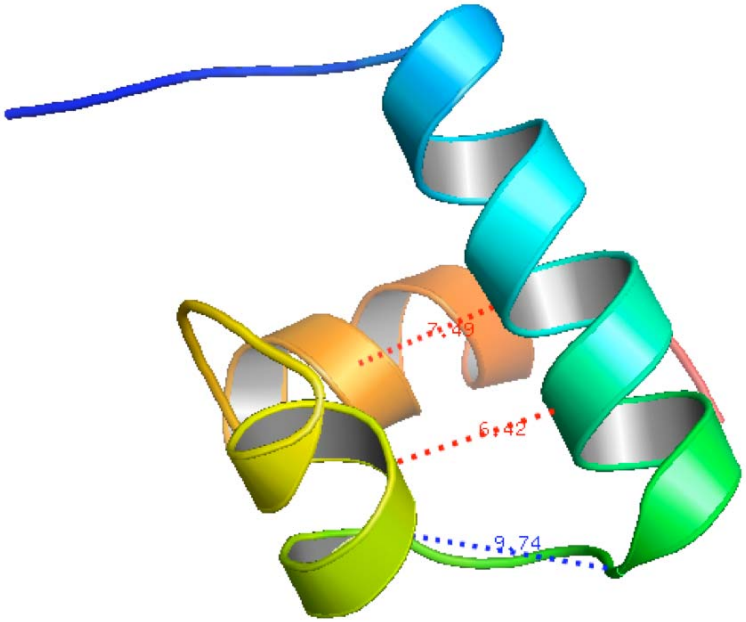
- **Level 2 Constraints**
  - **–Require at most one wheel contact for a specified primary contact**

$$\sum_k \sum_l w_{kl;ij}^{mn} \leq w_{ij}^{mn} \quad \forall (m,n,i,j) : y_{mn}^a = 1$$

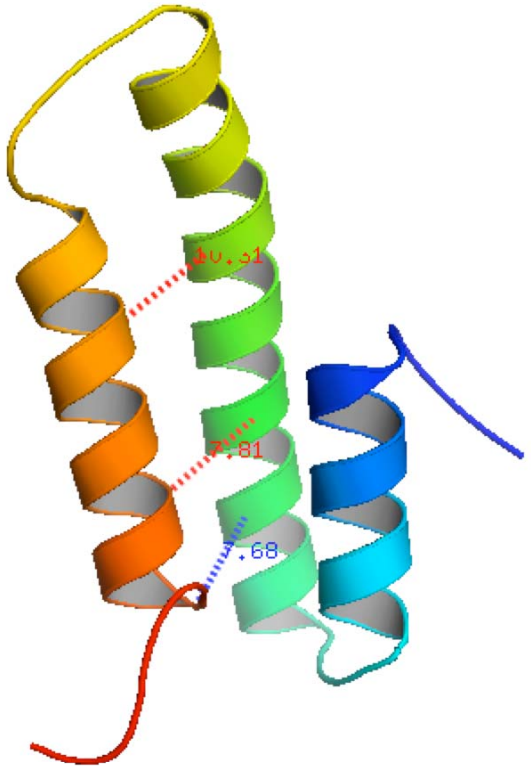
- **Level 2 Aim**
  - **–Distinguish between equally likely Level 1 predictions**
  - **–Increase the total number of contact predictions**

McAllister, Mickus, Klepeis, Floudas. Proteins. 2006, 65:930-952.

# Results – 2-3 helix bundles




PDB:1mbh in PyMol




PDB:1nre in PyMol

McAllister, Mickus, Klepeis, Floudas. Proteins. 2006, 65:930-952.

# Results – 1nre Contact Predictions

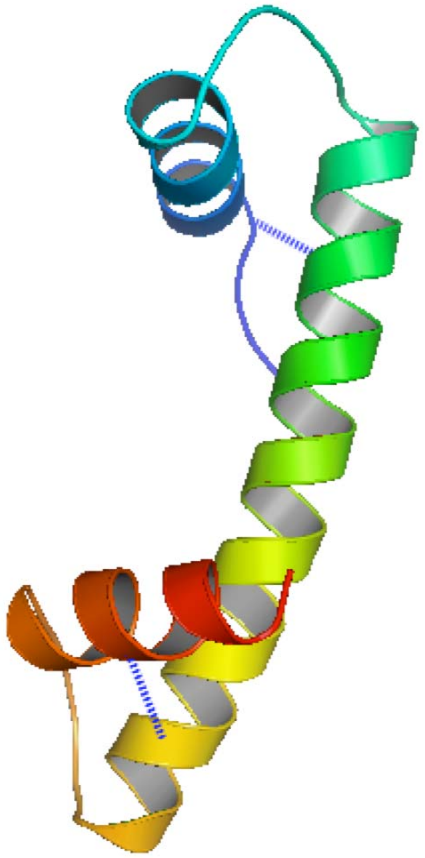
- subtract 0, max_contact 2

| PRIMARY Contact | PRIMARY Distance | WHEEL Contact | WHEEL Distance | Helix-Helix Interaction |
|---|---|---|---|---|
| 25L-49L | 6.0 | 28L-45L | 9.1 | 1-2 A |
| 28L-83V | 12.7 | - | - | 1-3 P |
| 45L-85L | 9.3 | 49L-81L | 8.1 | 2-3 A |
| 51I-77L | 9.3 | - | - | 2-3 A |

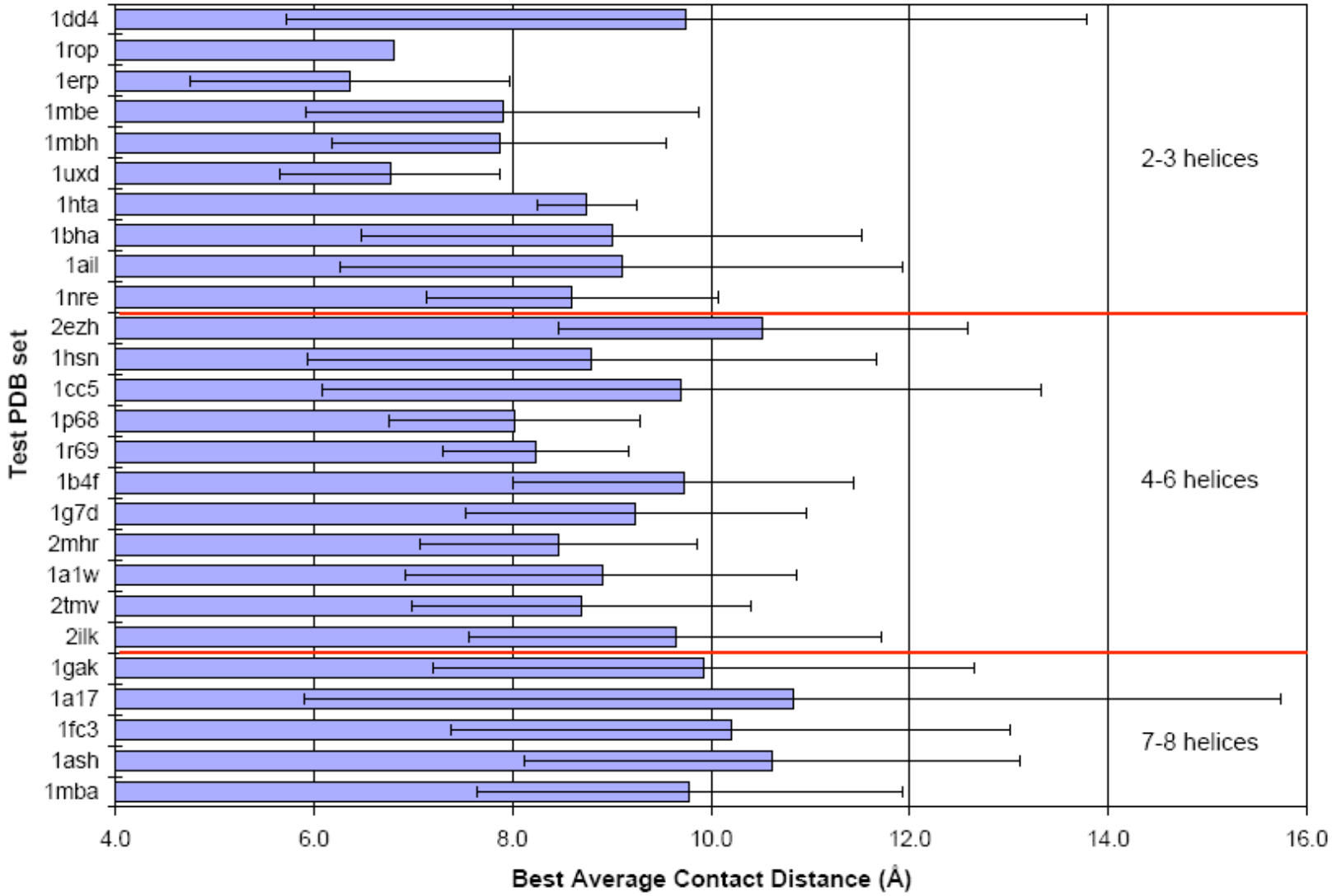
# Results – 1hta Contact Predictions

- subtract 0, max_contact 1

| PRIMARY Contact | PRIMARY Distance | Helix-Helix Interaction |
|---|---|---|
| 5I-28L | 9.1 | 1-2 A |
| 46L-62L | 8.4 | 2-3 A |



# Results – Contact Prediction Summary




McAllister, Mickus, Klepeis, Floudas. Proteins. 2006, 65:930-952.

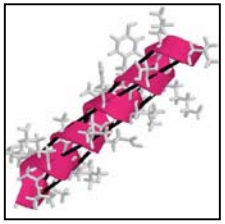
# Summary

- **Thesis: Topology of alpha helical globular proteins is based on inter-helical hydrophobic to hydrophobic contacts**
- Validated on alpha helical globular proteins

# Outline

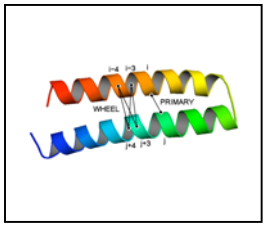
- Protein structure prediction overview
- Predicting α-helical contacts
  - Probability development
  - Model
  - Results
- Predicting α-helical contacts in α/β proteins
  - Distance bounding
  - Model
  - Results
- Structure prediction of α-helical proteins
  - Framework
  - Results

# ASTRO-FOLD for α-helical Bundles



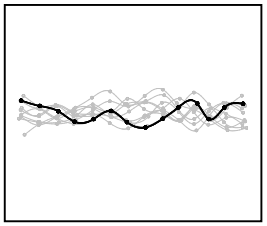
**Helix Prediction**
-Detailed atomistic modeling
-Simulations of local interactions

(Free Energy Calculations)




**Interhelical Contacts**
-Maximize common residue pairs
-Rank-order list of topologies

(MILP Optimization Model)

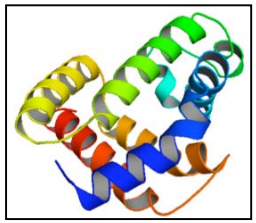


**Loop Structure Prediction**
-Dihedral angle sampling
-Discard conformers by clustering

(Novel Clustering Methodology)

**Derivation of Restraints**
-Dihedral angle restrictions
-$C^{\alpha}$ distance constraints

(Reduced Search Space)



**Overall 3D Structure Prediction**
-Structural data from previous stages
-Prediction via novel solution approach

(Global Optimization and Molecular Dynamics)

McAllister, Floudas. Proceedings, *BIOMAT 2005*.

# Hybrid Global Optimization Algorithm

- **All secondary nodes begin performing αBB iterations**
- Once the CSA bank is full, CSA takes control of a subset of secondary nodes

CSA work
•Rotamer optimization
•Minimization of CSA trial conformation

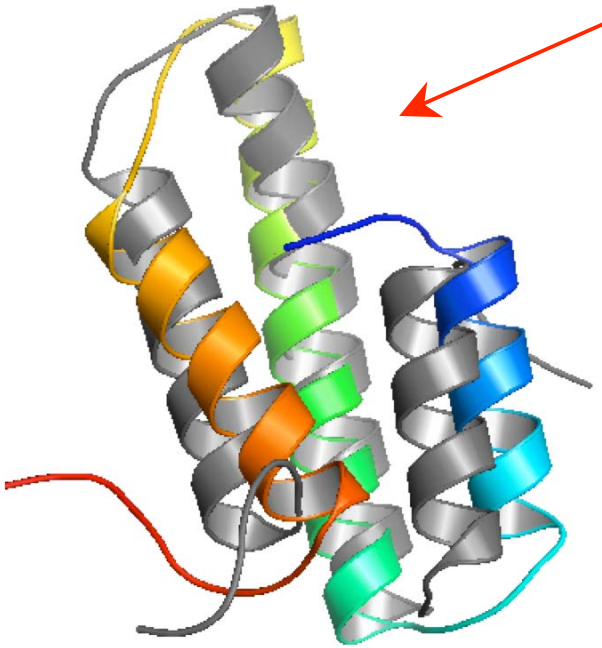
structures
•Only executed during idle time of primary processor

McAllister and Floudas. 2007, Submitted for publication.

# Results – Tertiary Structure Prediction

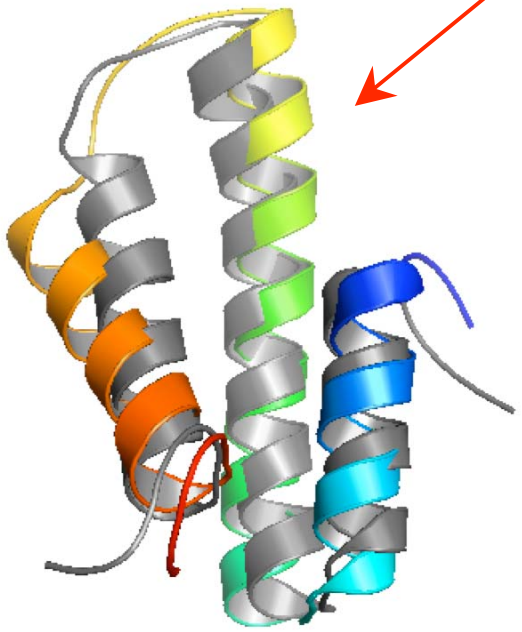
- PDB: 1nre

Energy -1395.48
RMSD 6.63



Lowest energy predicted structure of 1nre (color) versus native 1nre (gray)

Energy -1340.45
RMSD 3.52

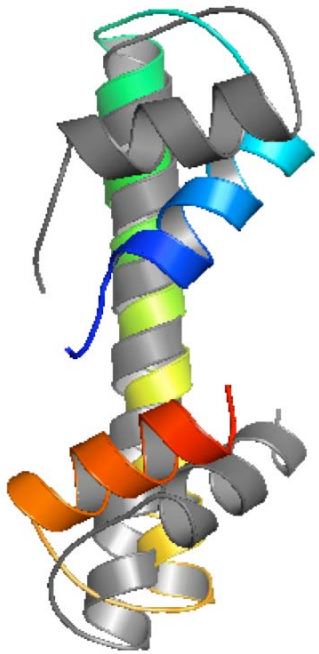


Lowest RMSD predicted structure of 1nre (color) versus native 1nre (gray)

# Results – Tertiary Structure Prediction

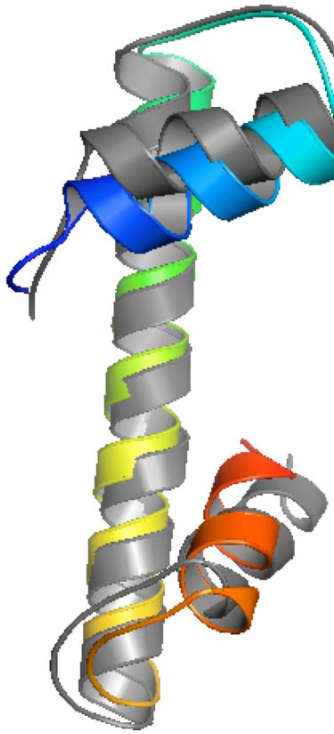
- PDB: 1hta

Energy -941.02
RMSD 6.70



Lowest energy predicted structure of 1hta (color) versus native 1hta (gray)

Energy -915.57
RMSD 2.58

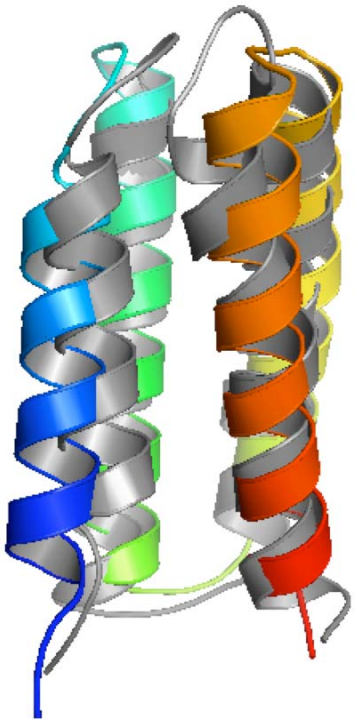


Lowest RMSD predicted structure of 1hta (color) versus native 1hta (gray)

# Results – Blind Tertiary Structure Prediction
## (Collaboration with Michael Hecht)

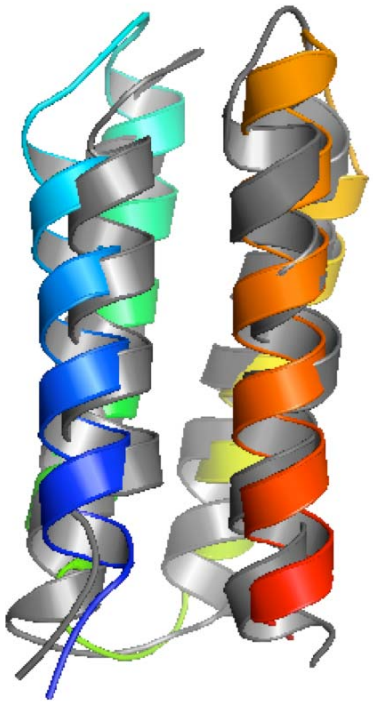
- S836

Energy -1740.11
RMSD 2.84



Lowest energy predicted structure of s836 (color) versus native s836 (gray)

Energy –1697.88
RMSD 2.39



Lowest RMSD predicted structure of s836 (color) versus native s836 (gray)

# Advances In *De Novo* Protein Design


Christodoulos A. Floudas

Princeton University

Department of Chemical Engineering

Program of Applied and Computational Mathematics

Department of Operations Research and Financial Engineering

Center for Quantitative Biology

# De Novo Protein Design

**Relevant References:**


- Klepeis J.L., C.A. Floudas, D. Morikis, C.G. Tsokos, E. Argyropoulos, L. Spruce, and J.D. Lambris, "Integrated Computational and Experimenal Approach for Lead Optimization and Design of Compstatin Variants with Improved Activity", Journal of the American Chemical Society, 125 (28), 8422-8423 (2003).
- Morikis D., A.M. Soulika, B. Mallik, J.L. Klepeis, C.A. Floudas, and J.D. Lambris, "Improvement of the anti-C3 activity of complement using rational and combinatorial approaches", Biochemical Society Transactions, 32, 28-32 (2003).
- Klepeis J.L., C.A. Floudas, D. Morikis, and J.D. Lambris, "Design of Peptide Analogs with Improved Activity using a Novel de novo Protein Design Approach", Industrial and Engineering Chemistry Research, 43, 3817-3826 (2004).
- Fung H.K., Rao S., Floudas C.A., Prokopyev O., Pardalos P.M., and F. Rendl, "Computational Comparison Studies of Quadratic Assignment Like Formulations for the In Silico Sequence Selection Problem in De Novo Protein Design", Journal of Combinatorial Optimization, 10, 41-60 (2005).
- Fung H.K., Taylor M.S. and C.A. Floudas, "Novel Formulations for the Sequence Selection Problem in de Novo Protein Design with Flexible Templates", Optimization Methods and Software, 22 (1), 51-71 (2007).
- Fung H.K., Floudas C.A, Taylor M.S., Zhang L., and D. Morikis, "Towards Full Sequence De Novo Protein Design with Flexible Templates for Human Beta-Defensin-2", Biophysical J., 94, 584-599 (2008).
- Taylor M.S., Fung H.K., Rajgaria R., Filizola M., Weinstein H., and C.A. Floudas, "Mutations Affecting the Oligomerization Interface of G-Protein Coupled Receptors Revealed by a Novel De Novo Protein Design Framework", Biophysical J., 94, 2470-2481 (2008).


**Review Articles**


- Floudas C.A., "Research Challenges, Opportunities and Synergism in Systems Engineering and Computational Biology", AIChE Journal, 51, 1872-1884 (2005).
- Floudas C.A., H.K. Fung, S.R. McAllister, M. Monnigmann, and R. Rajgaria, "Advances in Protein Structure Prediction and De Novo Protein Design: A Review", Chemical Engineering Science, 61, 966-988 (2006).
- Fung H.K., Welsh W.J., and C.A. Floudas, "Computational De Novo Peptide and Protein Design: Rigid Templates versus Flexible Templates", IECR, 47, 993-1001 (2008).

# 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 like framework
- Computational Studies
  - Design of Inhibitors for Complement 3
  - Design of C3a
- Conclusions & Acknowledgements

# Complement System

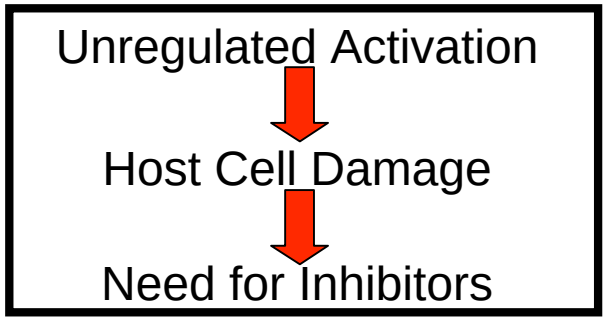
- ~30 distinct plasma proteins that interact to attack/eliminate pathogens
- Activated via (3) interacting pathways
  - (A) Classical : Antibody-binding (IgM, IgG) to pathogens
  - (B) Lectin : Mannose binding protein to carbohydrates on bacteria or viruses
  - (C) Alternative : Spontaneous binding to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

(A) : Adaptive/Acquired Immune Response

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

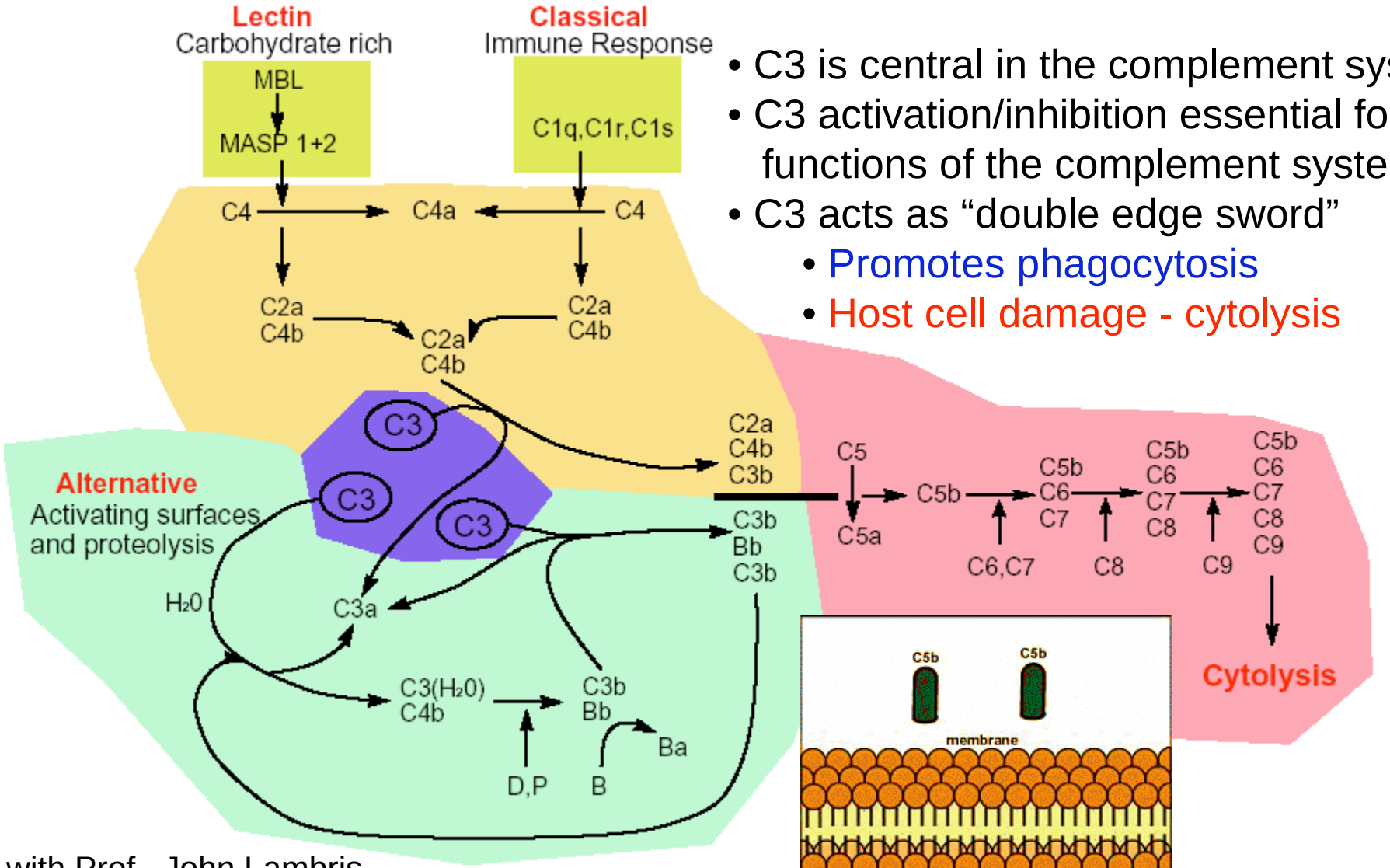
- Activation of the Complement System results in :
  - opsonization of pathogens (C3b; C4b)
  - recruitment of inflammatory cells (C3a; C5a; C4a+)
  - killing of pathogens (C5b, C6, C7, C8, C9)

MAC: Membrane Attack Complex




| | Annual US Patient Population |
|---|---|
| • Acute Complement-Mediated Conditions | |
| • 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

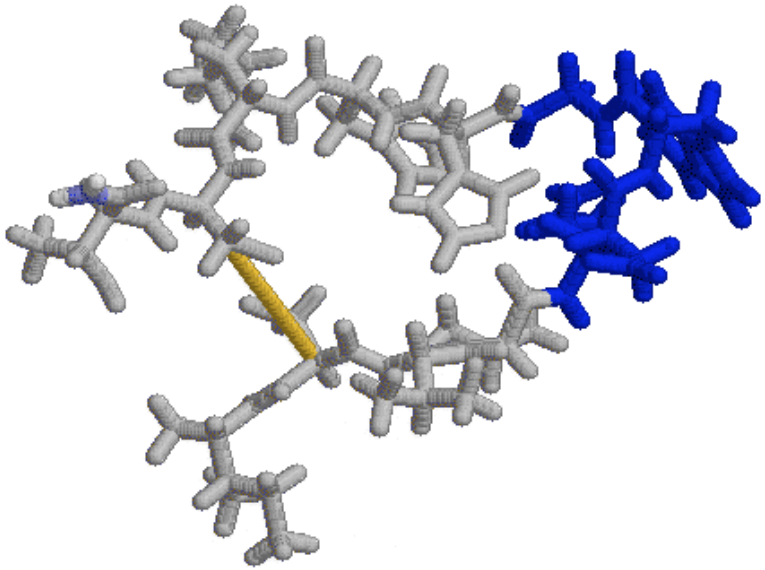
With Prof. John Lambris, University of Pennsylvania, School of Medicine
With Prof. Dimitris Morikis, University of California at Riverside

## 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

## A potent mediator of inflammation

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

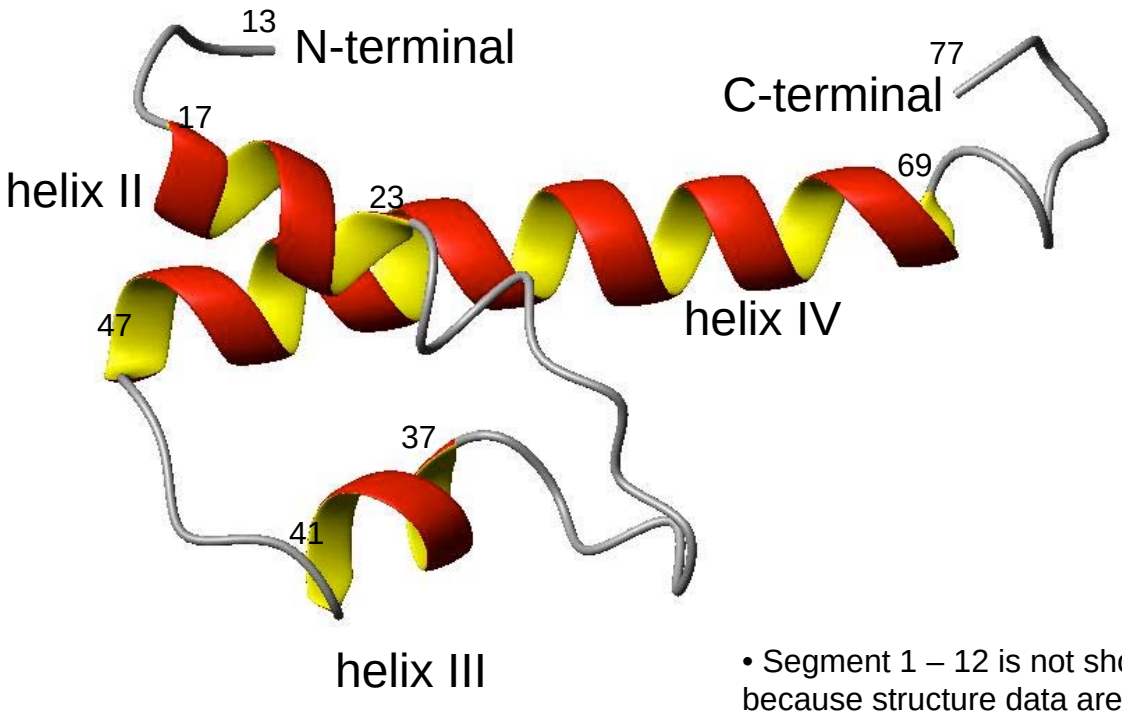
### Background

- 77 residues, 3 S-S bonds, 4 $\alpha$-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 positions 1-12 not resolved

# Structure of C3a

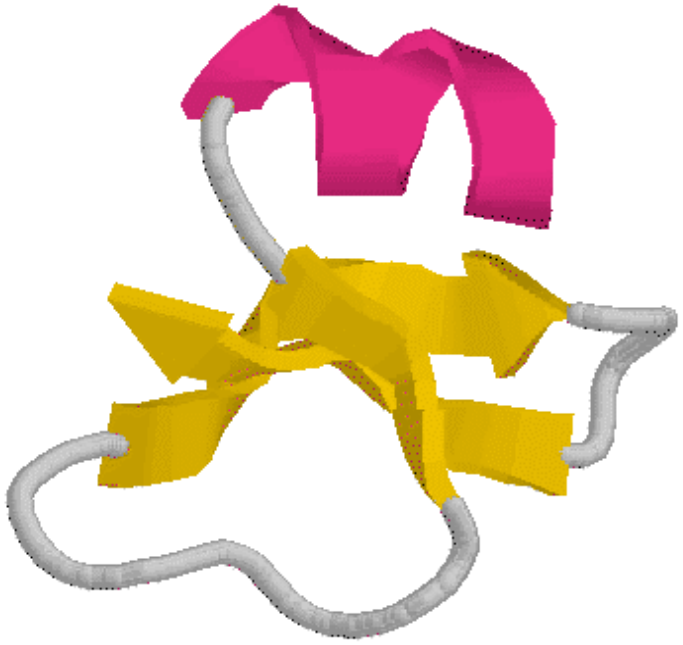



- Segment 1 – 12 is not shown because structure data are not available.
- helix I is segment 5 – 15.

# 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 |

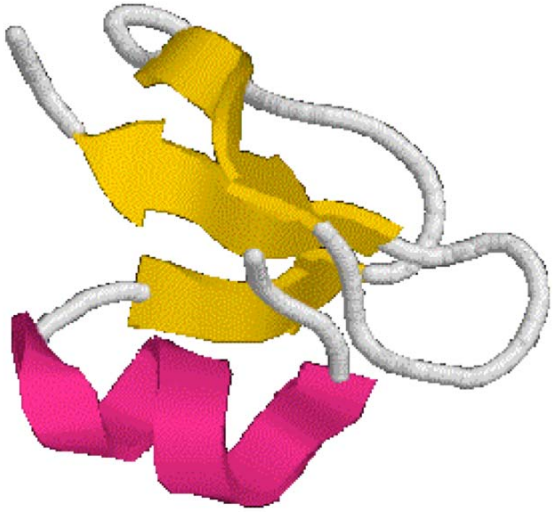
## Objective

Design improved antibacterial peptides

# De Novo Protein Design

## Define target template

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

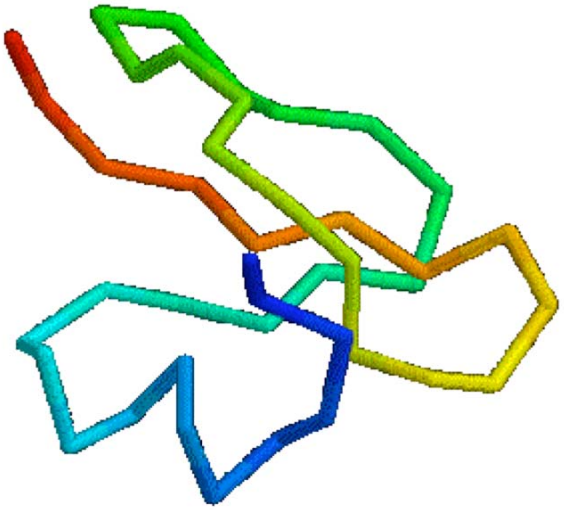


Human β-Defensin-2

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.

## Challenges

In silico sequence selection

Fold specificity

Combinatorial complexity

-Backbone length : n

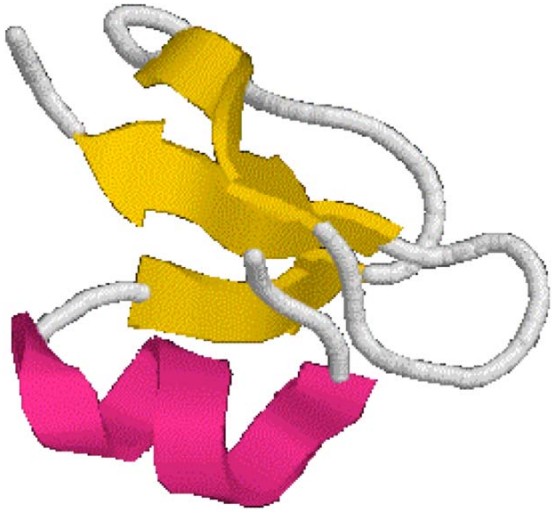
-Amino acids per position : m

$m^n$ possible sequences

# De Novo Protein Design

## Define target template

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

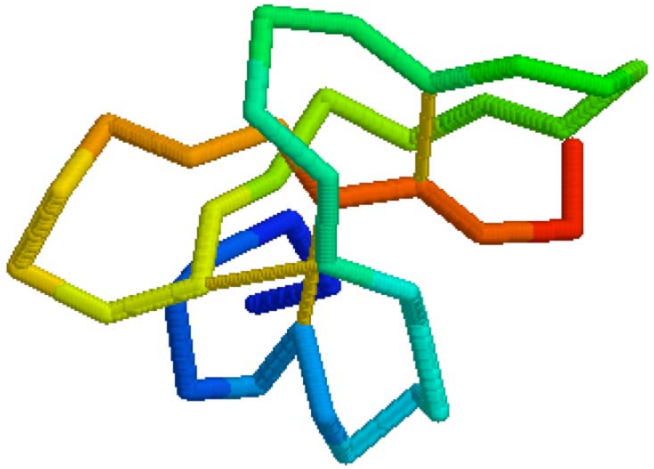


Human β-Defensin-2

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.

## Challenges

In silico sequence selection

Fold validation/specificity

Combinatorial complexity

-Backbone length : n

-Amino acids per position : m

$m^n$ possible sequences

# 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

# Background and Advances

- Stochastic Methods: MC, Genetic Algorithms

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

- Probabilistic Approaches, Combinatorial Libraries

  Saven & co-workers (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 Criterion (Mayo & coworkers;Desjarlais, Handel,1995,1999; Hellinga & co-workers; Desmet at al. 1992; Goldstein, 1994;Pierce et al. 2000)

- Iterative Sequence-Structure (Kuhlman et al. 2003)

# Background

## Different De Novo Protein Design Approaches

- Stochastic methods:
  - Monte Carlo Metropolis *et al*., J. Chem. Phy., 1953

    - Perturb the structure by some random change in residue or rotamer. Move is accepted if Boltzmann probability is higher than some random number.

  - Genetic algorithms Tuffery *et al*., J. Biomol. Struct. Dyn., 1991

    -Random sequences are allowed to mutate, cross-over, and reproduce. High energy sequences are eliminated from population.

  - Stochastic methods do not guarantee convergence to the global energy minimum

# Background

## Different De Novo Protein Design Approaches

- Combinatorial Libraries: Probabilistic approach

Zou and Saven, J. Mol. Bio., 2000
Kono and Saven, J. Mol. Bio., 2001
Park et al., Current Opinion in Structural Biology, 2004

- set of 20 probabilities for each design position

- maximize the total conformational entropy subject to constraints:

$$\max \quad -\sum_{i,\alpha,r(\alpha)} w_i(\alpha, r(\alpha)) \ln w_i(\alpha, r(\alpha)) - \lambda\left(\sum_{\alpha,r(\alpha)} w_i(\alpha, r(\alpha)) - 1\right) - \beta(E - E_o)$$

*i*: position  α: amino acid  *r*(α): rotamer

Lagrange multipliers

- provide framework for designing and interpreting protein combinatorial experiments

# Background

## Different De Novo Protein Design Approaches

- Deterministic methods:
  - Self-Consistent Mean Field — Koehl and Delarue, J. Mol. Bio., 1994; Lee, J. Mol. Bio., 1994

    - refines iteratively a conformational matrix

    - initial guess for conformational matrix: $CM(i,k) = \frac{1}{K_i}$ for rotamer *k* of residue *I*

    - mean field energy: $E(i,k) = U(x_{ikC}) + U(x_{ikC}, x_{0C}) + \sum_{j=1, j\neq i}^{N} \sum_{l=1}^{K_j} CM(j,l) U(x_{ikC}, x_{jlC})$

      - $U(x_{ikC})$: Coordinates of atoms of residue *i* described by rotamer *k*
      - $U(x_{ikC}, x_{0C})$: Coordinates of atoms in template

    - update conformational matrix: $CM_1(i,k) = \frac{e^{\frac{-E(i,k)}{RT}}}{\sum_{l=1}^{K_i} e^{\frac{-E(i,l)}{RT}}}$

      $CM = \lambda CM_1 + (1-\lambda) CM$

      optimal $\lambda = 0.9$

    - convergence criterion (e.g., $10^{-4}$) is set to define self-consistency

# Background

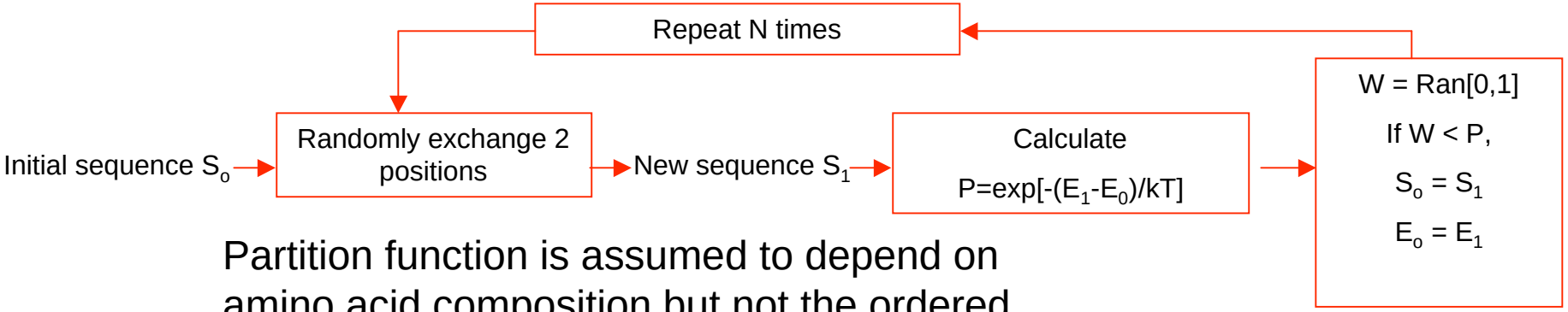
## Different De Novo Protein Design Approaches

- Self-Consistent Mean Field + Monte Carlo

Koehl and Levitt, J. Mol. Bio., 1999I,II

- energy function: full-atomistic using self-consistent mean field approach

- "Design in" for stability and "design out" for specificity, using the approximation of a random energy model:




Partition function is assumed to depend on amino acid composition but not the ordered sequence → specificity (confirmed by fold recognition techniques) achieved by optimizing sequence space and holding amino acid composition fixed

# Background

## Different De Novo Protein Design Approaches

- Deterministic methods:
  - Dead-End Elimination

Desmet *et al.*, Nature, 1992
Voigt *et al.*, J. Mol. Bio., 2000
Pierce *et al.*, J. Comput. Chem., 2000
Gordon *et al.*, J. Comput. Chem., 2003

- systematically eliminate rotamers that are incompatible with the lowest energy sequence

- energy function: $$E = \sum_{i=1}^{N} E(i_r) + \sum_{i=1}^{N-1} \sum_{j>i}^{N} E(i_r, j_s)$$

Rotamer-template | Rotamer-rotamer

- different dead-end elimination criteria:

$$E(i_r) + \sum_{j \neq i}^{N} \min_s E(i_r, j_s) > E(i_t) + \sum_{j \neq i}^{N} \max_s E(i_t, j_s)$$ Original DEE

$$E(i_r) - E(i_t) + \sum_{j \neq i}^{N} \min_s [E(i_r, j_s) - E(i_t, j_s)] > 0$$ Simple Goldstein DEE

$$E(i_r) - E(i_t) + \sum_{j, j \neq k \neq i}^{N} \{\min_u [E(i_r, j_u) - E(i_t, j_u)]\} + [E(i_r, k_v) - E(i_t, k_v)] > 0$$ Simple / Split DEE

# Background

## Different De Novo Protein Design Approaches

- Dead-End Elimination

$$E(c) - E(c') + \min_{\{j_s\} \| c \cup \{j_s\}} \sum_{j=1, \notin r}^{p} [E(c, j_s) - E(c', j_s)] > 0$$

Generalized DEE criterion

*c*: query cluster

*c*': comparison cluster

Looger & Hellinga, J. Mol. Bio., 2001

## Fundamental Assumptions of Dead-End Elimination

Fixed backbone template

Discrete set of rotamers

# Background

- Deterministic methods guarantee convergence to the global minimum

- Self-Consistent Mean Field and Dead-End Elimination methods either:

1. assume a fixed template
2. fix the amino acid composition
2. consider an average set of templates
3. consider a discrete set of rotamers

True backbone flexibility is not allowed

# De Novo Protein Design: Advances

## Conferring novel functions onto template

- DEZYMER program for designing metalloproteins

1. First identify the catalytic functional groups that catalyze the desired reaction
2. Relocate these groups from mother sequence to the best positions in the de novo designed protein

Richards and Hellinga, J. Mol. Bio., 1991

Richards *et al*., J. Mol. Bio., 1991

Benson *et al*., Proc. Nat. Acad. Sci. USA, 2000

- Succeeded to create Zn, FeS, and Cu binding sites in thioredoxin, a protein which normally does not bind metals (Richards *et al*., J. Mol. Bio., 1991)

# De Novo Protein Design: Advances

## Better stability and specificity

- Engineered $\alpha$-lytic protease showed over 200-fold preference for one substrate kind over another

Wilson *et al*., J. Mol. Bio., 1991

- Redesigned compstatin (complement 3 inhibitor) found to have the best inhibitory activity of 16-fold more potent than the parent peptide

Klepeis *et al*., *J. Am. Chem. Soc*., 2003

Klepeis *et al*., *Ind. & Eng. Chem. Research*, 2004

- Higher stability by having more hydrophobic amino acids in the core than parent proteins

Kuhlman & Baker,

Current Opinion in Structural Biology , 2004

- Redesign protein-protein interfaces

Kortemme & Baker,

Current Opinion in Chem. Bio., 2004

# De Novo Protein Design: Advances

## Locking proteins into particular conformations

- Enforced integrin I, a cell-surface adhesion receptor that binds with complement component iC3b, to adopt either the open or closed conformation

Shimaoka *et al*., Nat. Struct. Bio., 2000

- Restricted amino-terminal domain of calmodulin to its calcium saturated closed form.

Kraemer-Pecore *et al*.,
Current Opinion in Chem. Bio., 2001

many more other successes, but…

# Flexibility of Backbone Templates

- Scaling down the atomic van der Waals radii by a factor (~5-10%)

  **Dahiyat, B.I. & Mayo, S.L. (1997) *Proc. Natl. Acad. Sci. USA*, 94, 10172-10177.**

  **- Overestimation of attractive forces between atoms and the possibility of atom overpacking**

- Considering a fixed set of rotamers (DEE) or changing super-secondary structure parameters which alter relative orientation and distance between secondary structures

  **Su, A. & Mayo, S.L. (1997) , *Protein Science*, 6, 1701-1707.**

  **- Only a subset of possible conformations is considered**

- Generating ensemble of random structures from template

  **Desjarlais, J.R. & Handel, T.M. (1999),*J. Mol. Bio.*, 289, 305-318.**

  **- 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**

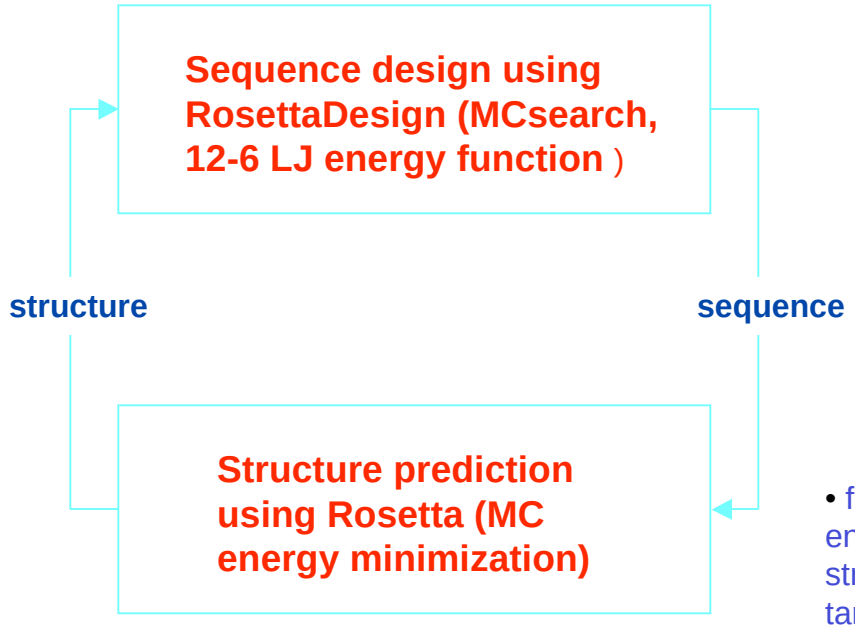
  **- Only a random subset of possible conformations is considered**

# Flexibility of Backbone Templates

- Iterating between sequence space and structure space

Saunders, C.T. & Baker D. (2005) *J. Mol. Bio.*, 346, 631-644.

Kuhlman, B. & Dantas G. & Ireton G.C. & Varani G. & Stoddard B.L. & Baker D. (2003) *J. Mol. Bio.*, 302, 1364-1368.




• for each of an ensemble of starting structures around the target fold

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

# 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°.

- $C^{\alpha}$-$C^{\alpha}$ 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**

# 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, Taylor, Floudas, *OMS*, 2007)

- Distance-dependent $C^{\alpha}$-$C^{\alpha}$, centroid-centroid forcefields

(Loose, Klepeis, Floudas, *Proteins* (2004); Rajgaria, McAllister, Floudas, *Proteins*, 2006; Rajgaria, McAllister, Floudas, *Proteins*, Accepted, 2007)

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

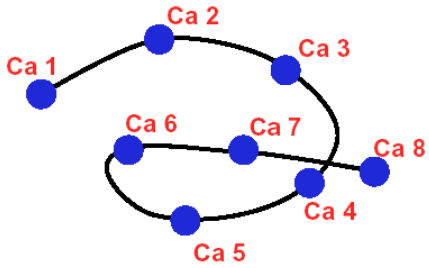
- First principles via ASTRO-FOLD

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

- NMR structures refinement-based method via CYANA and TINKER using AMBER

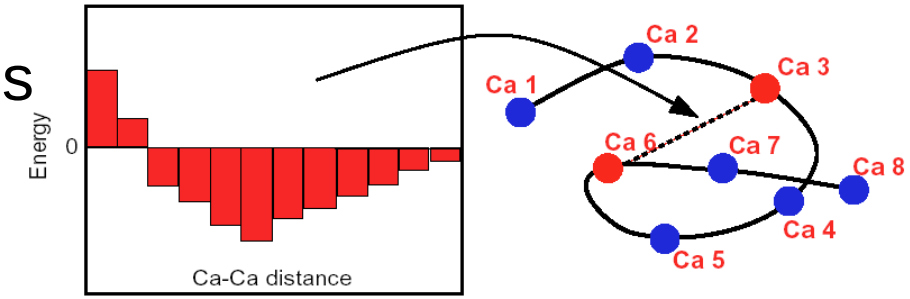
# De Novo Protein Design Framework

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




## Sequence selection

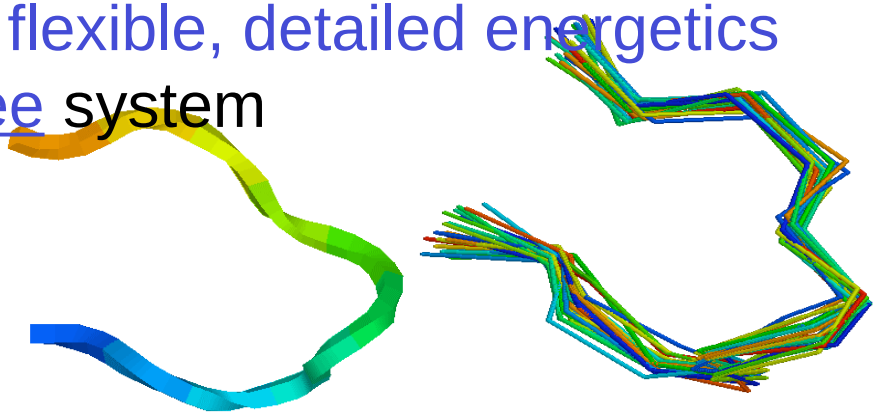
- Identify target template for desired fold; specify coordinates of backbone
- Identify possible residue mutations
- ***Distance dependent pairwise potentials***
- Generate rank-ordered energetic list from mixed-integer linear (MILP)




| 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)

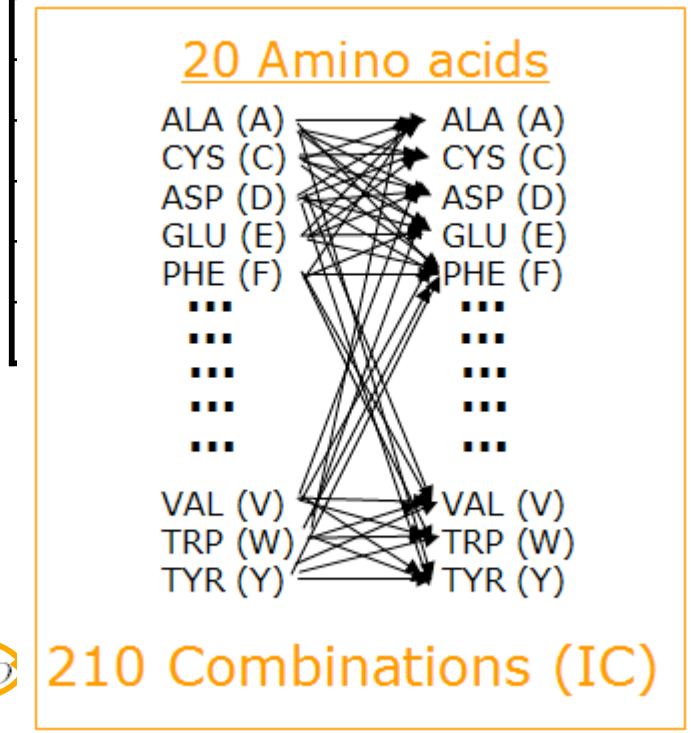
More resolution for bin 3 to 6

210 amino-acid combination (IC)

Table 1: Bin Definition

| Bin ID | $C^{\alpha}$-distance [$A^{o}$] |
|---|---|
| **1** | **3-4** |
| **2** | **4-5** |

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




## 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

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

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

$$\theta_{IC,ID+1} - \theta_{IC,ID} \geq -4, \quad \forall IC; ID = 2, 3, ...7$$

$$\theta_{IC,ID+1} - \theta_{IC,ID} \leq 4, \quad \forall IC; ID = 2, 3, ...7$$

Smooth profile (Tobi and Elber*, 2000)

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

$$\theta_{IC,ID} \geq -4, \quad \forall IC; ID = 8$$

$$\theta_{IC,ID} \leq 4, \quad \forall IC; ID = 8$$

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

# 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")

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

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

# Hydrophobic Constraints*

Captures interaction between certain amino acids (Bahar and Jernigan, 1997)

Favorable interaction at 4-6.5 Å between hydrophobic groups

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

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

| Hydrophilic (Neut) $\{\mathcal{PU}\}$ | Hydrophilic (Pos) $\{\mathcal{PP}\}$ | Hydrophobic Non-Aromatic $\{\mathcal{HN}\}$ | Hydrophobic Aromatic $\{\mathcal{HA}\}$ |
|---|---|---|---|
| GLY | LYS | CYS | PHE |
| HIS | ARG | ILE | TYR |
| ASN | Hydrophobic (Neg) $\{\mathcal{PN}\}$ | LEU | TRP |
| PRO | | MET | Other $\{\mathcal{O}\}$ |
| GLN | ASP | THR | |
| SER | GLU | VAL | ALA |

Table 2: Amino Acid Classification

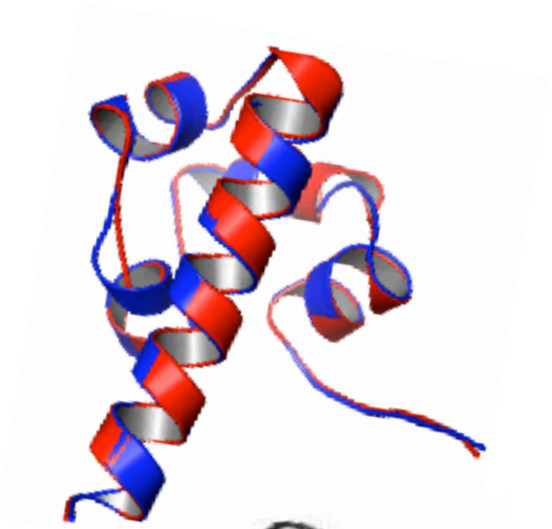
$$\theta_{IC,ID+1} - \theta_{IC,ID} \leq -4, \quad IC \in \{H,H\}; ID = 1$$
$$\theta_{IC,ID+1} - \theta_{IC,ID} \leq -2, \quad IC \in \{H,H\}; ID = 2,3$$
$$\theta_{IC,ID+2} - \theta_{IC,ID} \geq 0, \quad IC \in \{H,H\}; ID = 4$$
$$\theta_{IC,ID+1} - \theta_{IC,ID} \leq 2, \quad IC \in \{H,H\}; ID = 4$$

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

# High Resolution Decoys* - Idea

## Goal

To create a large number of near-native protein structures for a non-homologous set of proteins that span the Protein Data Bank.



## Hypothesis

High quality near-native structures maintain similar $C^{\alpha}$-$C^{\alpha}$ distances for the hydrophobic residues contained in the elements of secondary structure.

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

# 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 Ca-Ca Distance Dependent 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 Centroid based Force Field

*Rajgaria. R., McAllister. S. R., and Floudas C.A.

# Side Chain Based Force Field

## Importance

– $C^{\alpha}$ 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

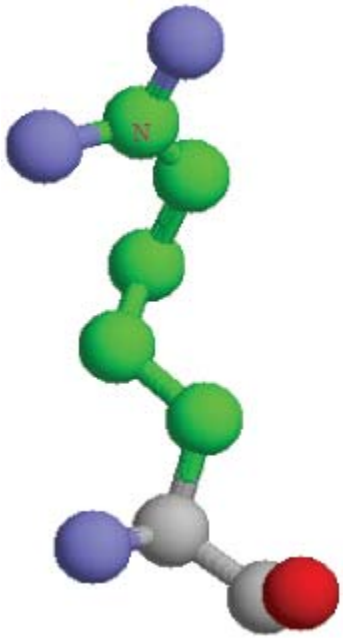
Need to revisit the interaction center definition

"effective" distance range might be different

define "centroid"

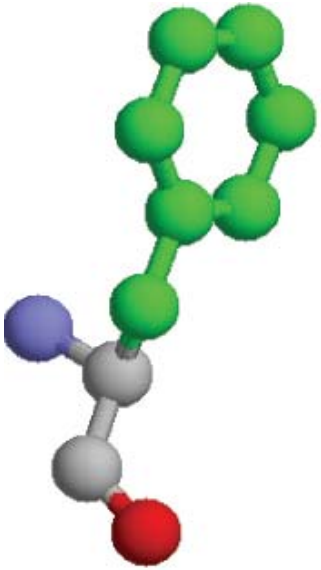
# Force Field – Side Chain Based Formulation

Side Chain Centroid definition

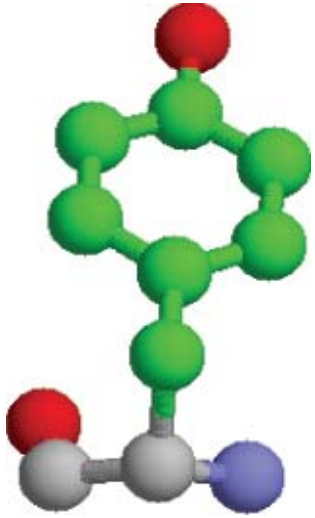


Arg

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

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

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

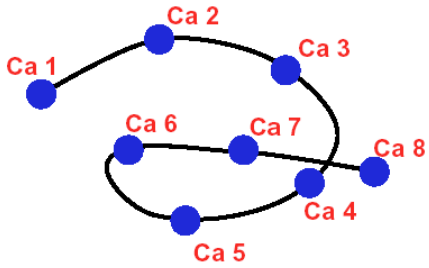
| FF name | Avg. Rank | # Firsts | Avg.RMSD (Å) | Avg. Z-score |
|---|---|---|---|---|
| 6bin-HRSC | 2.49 | 128/148 | 0.29 | 3.62 |
| 7bin-HRSC | 2.01 | 125/148 | 0.32 | 3.39 |
| HR | 1.87 | 113/150 | 0.45 | 2.11 |
| LKF | 39.45 | 17/150 | 1.72 | 1.55 |
| TE-13* | 19.94 | 92/148 | 0.81 | 3.15 |

***Rajgaria. R., McAllister. S. R., and Floudas C.A. , Proteins, Accepted for publicaion, (2007)***

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

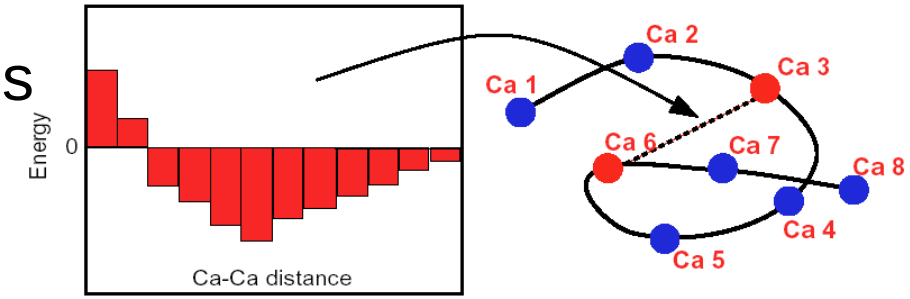
# De Novo Protein Design Framework

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




## Sequence selection

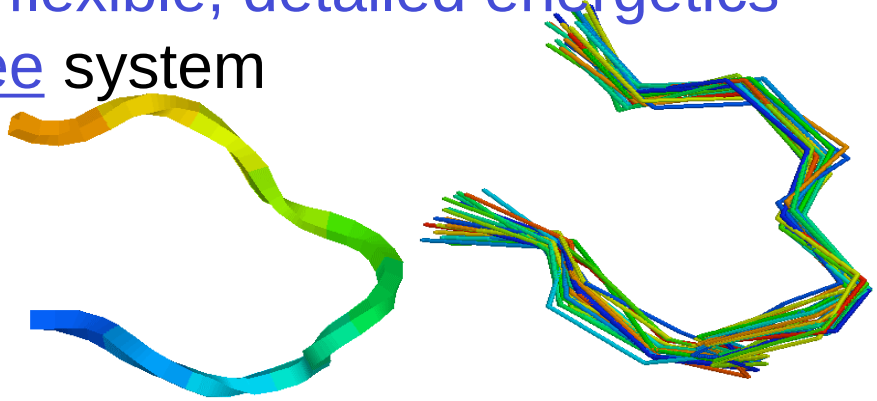
- Identify target template for desired fold; specify coordinates of backbone
- Identify possible residue mutations
- Introduce distance dependent pairwise potential based on Ca
- Generate ***rank-ordered list from mixed-integer linear (MILP)***




| 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 $i = 1,2,\ldots,n$ there can be $j = 1,2,\ldots,m_i$ mutations
- Define equivalent sets $k = 1,2,\ldots,n$ and $l = 1,2,\ldots,m_k$
- Require $k > i$ to represent all unique interactions
- Introduce 0-1 variables to indicate possible mutations at a given position

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

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

# Mixed-integer Nonlinear Model

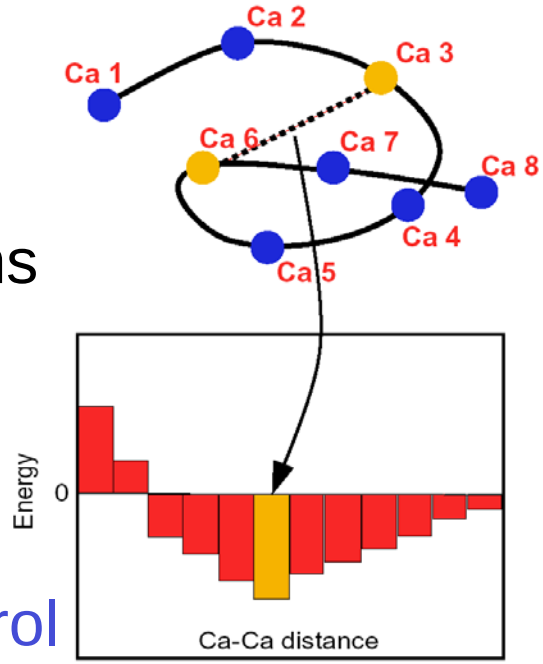
$$\min E(x) = \sum_{i=1}^{n} \sum_{j=1}^{m_i} \sum_{k>i}^{n} \sum_{l=1}^{m_k} E_{ik}^{jl}(x_i^j, x_k^l) y_i^j y_k^l$$

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

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

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

- $E_{ik}^{jl}$ 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

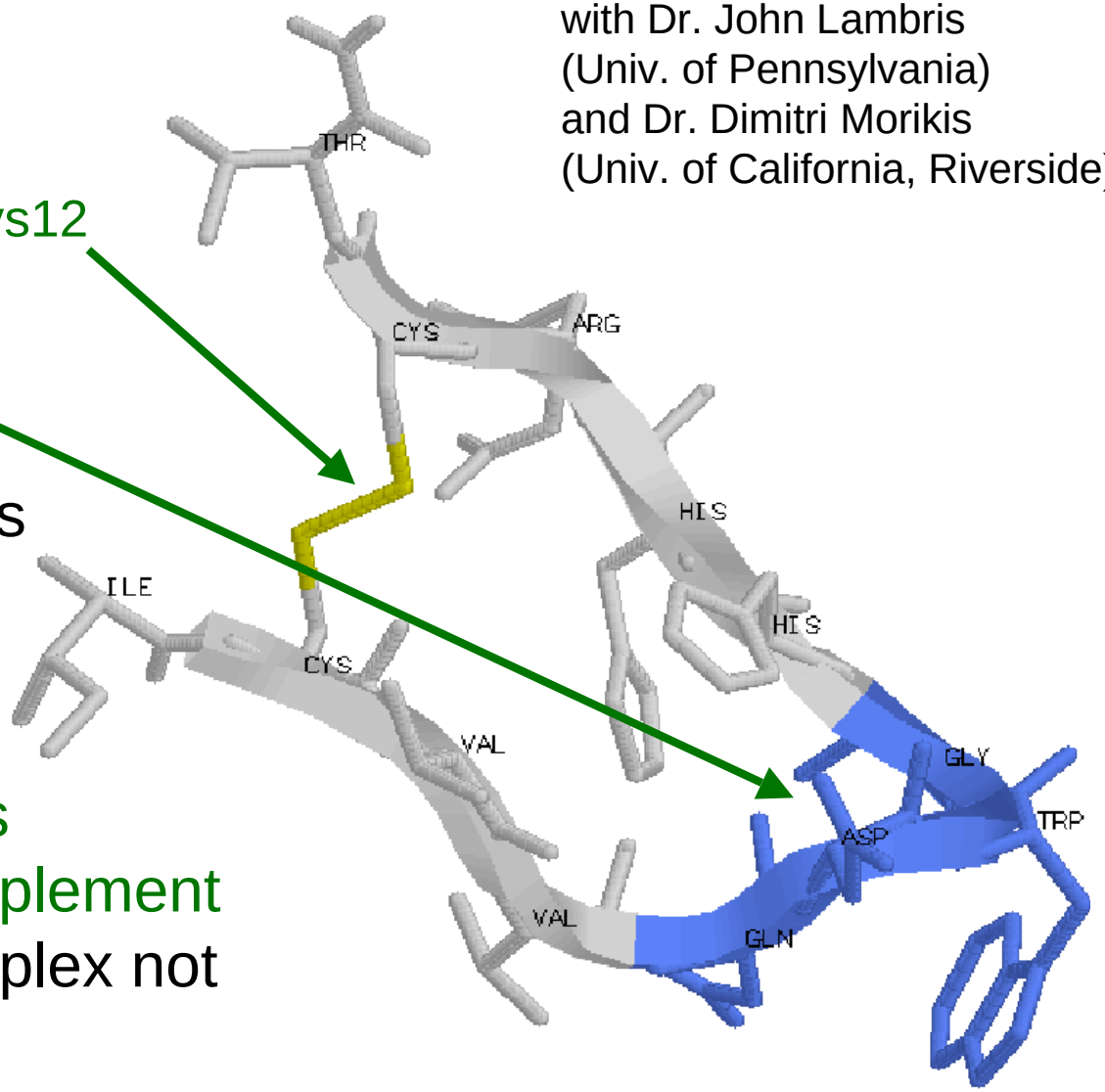
# Compstatin

## 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 |
| 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)

# 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:

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

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

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

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

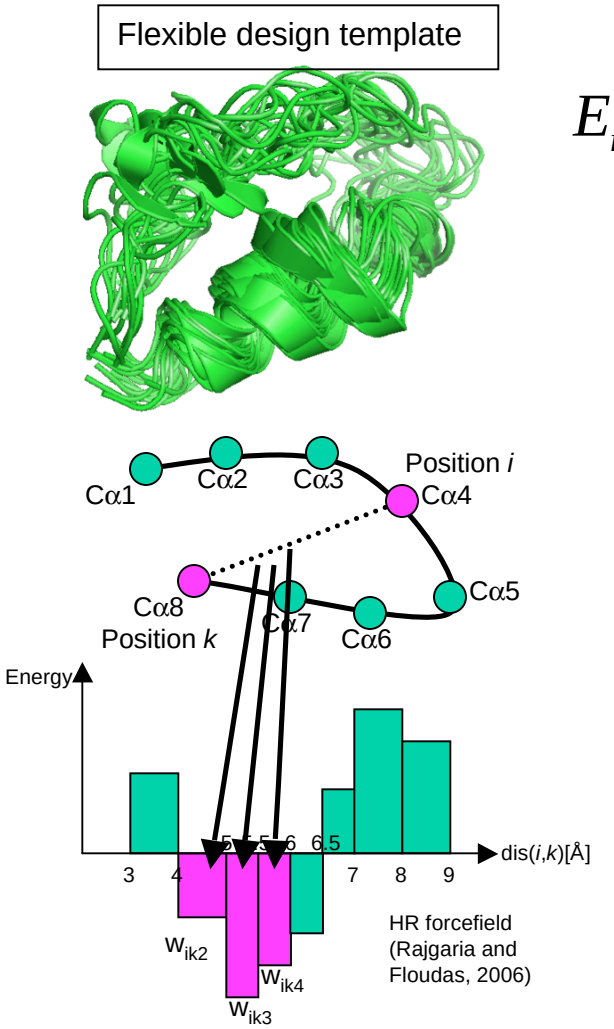
$$y_i^j, y_k^l, w_{ik}^{jl} = 0-1 \qquad \forall i, j, k > i, l$$

- 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




$$E_{ik}^{jl} \qquad \sum_{n} w_{ikn} E_{ik}^{jl}$$ 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

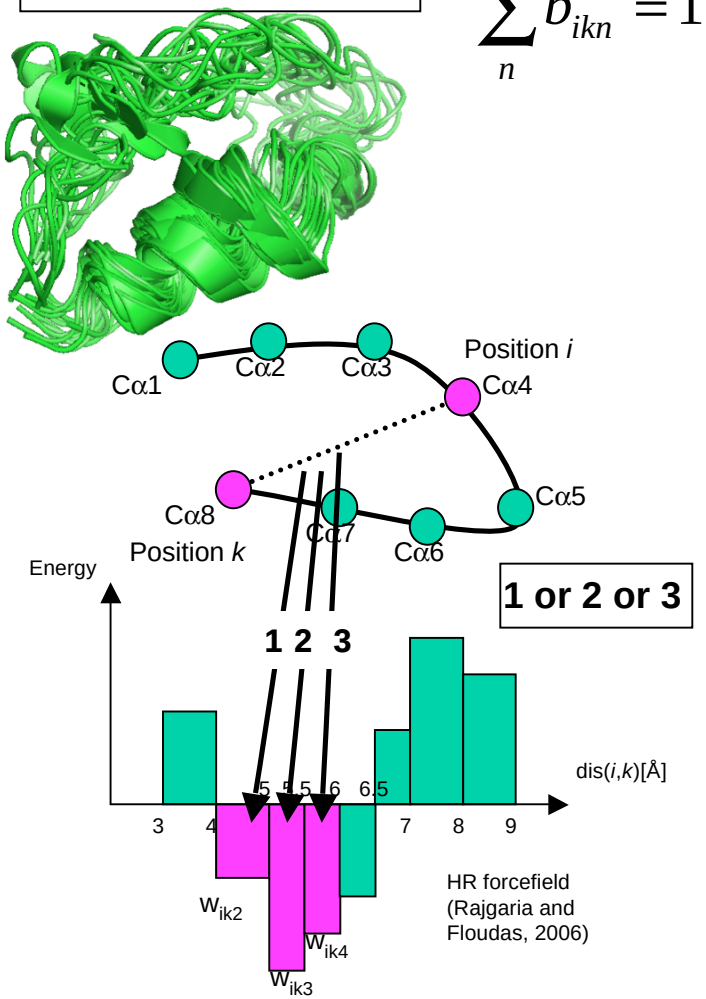
$E_{ik}^{jl}$ changed to $\sum_n b_{ikn} E_{ik}^{jl}$ in objective function

$\sum_n b_{ikn} = 1$

$b_{ikn}$ = 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 $disbin(x_k,x_p,d')=1$

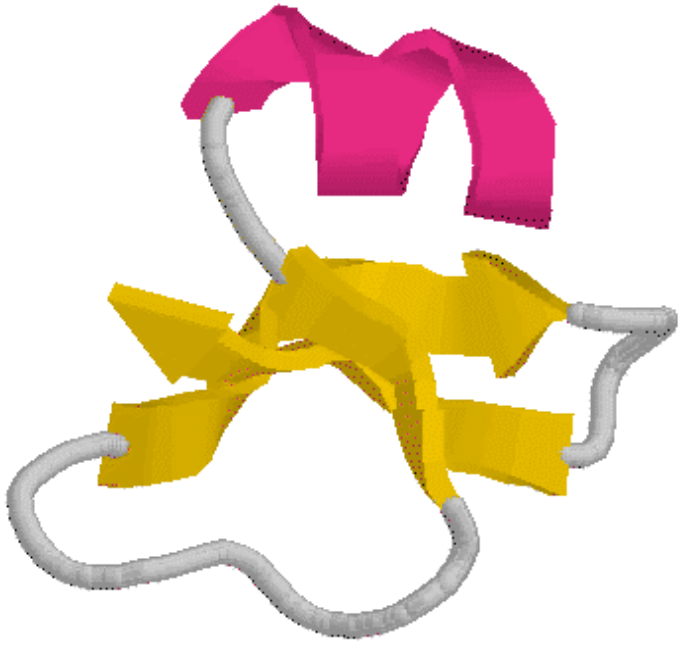
$\forall i,k>i,p,d,d',i \ne k \ne p$

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

# 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<br>25 - 28<br>36 - 39 |
| α Helix | 5 - 10 |
| S-S bonds | 8 - 37<br>15 - 30<br>20 - 38 |
| β-Turns | 16 - 19<br>21 - 24<br>32 - 35 |
| Hairpins | 25 - 29 |
| Bulges | 27, 28, 37 |

Constraints to add to model:

At least 2 hydrophobics on each β strand:

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

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

# De Novo Design of hβD-2

## More constraints to add…

Total number of hydrophobics more than wild type sequence for higher stability (Kuhlman & Baker, Cur. Op. Struct. Bio., 2004):

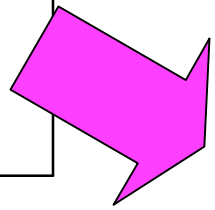
$$\sum_{i,j}\left(y_i^{Ala}+y_i^{Cys}+y_i^{Ile}+y_i^{Leu}+y_i^{Met}+y_i^{Phe}+y_i^{Trp}+y_i^{Tyr}+y_i^{Val}\right)\geq 17 \qquad \forall i$$

- Using PSI-BLAST, homology search was run to determine properties that are conserved among hβD and similar sequences.

- Conserved properties are translated into constraints:

Charge properties of the 97 hβD homologs from PSI-BLAST

| | lower bound | upper bound |
|---|---|---|
| Positive charges | 5 | 10 |
| Negative charges | 0 | 2 |
| Net charges | 4 | 9 |
| Total helix charge | 0 | +3 |



$$5\leq\sum_{i,j}\left(y_i^{Arg}+y_i^{Lys}\right)\leq 10 \qquad \forall i$$

$$0\leq\sum_{i,j}\left(y_i^{Asp}+y_i^{Glu}\right)\leq 2 \qquad \forall i$$

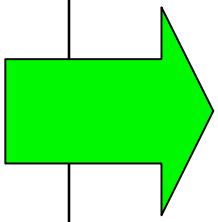
$$4\leq\sum_{i,j}\left(y_i^{Arg}+y_i^{Lys}-y_i^{Asp}-y_i^{Glu}\right)\leq 9 \qquad \forall i$$

$$0\leq\sum_{i,j}\left(y_i^{Arg}+y_i^{Lys}-y_i^{Asp}-y_i^{Glu}\right)\leq 3 \qquad \forall 5\leq i\leq 10$$

# De Novo Design of hβD-2

## More constraints to add…

| Amino acid | lower bound | upper bound |
|---|---|---|
| Ala | 0 | 3 |
| Arg | 1 | 9 |
| Asn | 0 | 6 |
| Asp | 0 | 2 |
| Cys | 4 | 7 |
| Gln | 0 | 3 |
| Glu | 0 | 3 |
| Gly | 3 | 7 |
| His | 0 | 4 |
| Ile | 0 | 6 |
| Leu | 0 | 4 |
| Lys | 0 | 7 |
| Met | 0 | 3 |
| Phe | 0 | 4 |
| Pro | 0 | 5 |
| Ser | 0 | 6 |
| Thr | 0 | 4 |
| Trp | 0 | 1 |
| Tyr | 0 | 4 |
| Val | 0 | 6 |



$$0 \le \sum_{i,j} y_i^{Ala} \le 3 \quad \forall i \qquad 1 \le \sum_{i,j} y_i^{Arg} \le 9 \quad \forall i$$

$$0 \le \sum_{i,j} y_i^{Asn} \le 6 \quad \forall i \qquad 0 \le \sum_{i,j} y_i^{Asp} \le 2 \quad \forall i$$

$$6 \le \sum_{i,j} y_i^{Cys} \le 6 \quad \forall i \qquad 0 \le \sum_{i,j} y_i^{Gln} \le 3 \quad \forall i$$

$$0 \le \sum_{i,j} y_i^{Glu} \le 3 \quad \forall i$$

$$0 \le \sum_{i,j} y_i^{His} \le 4 \quad \forall i \qquad 0 \le \sum_{i,j} y_i^{Ile} \le 6 \quad \forall i$$

$$0 \le \sum_{i,j} y_i^{Leu} \le 4 \quad \forall i \qquad 0 \le \sum_{i,j} y_i^{Lys} \le 7 \quad \forall i$$

$$0 \le \sum_{i,j} y_i^{Met} \le 3 \quad \forall i \qquad 0 \le \sum_{i,j} y_i^{Phe} \le 4 \quad \forall i$$

$$5 \le \sum_{i,j} y_i^{Pro} \le 5 \quad \forall i \qquad 0 \le \sum_{i,j} y_i^{Ser} \le 6 \quad \forall i$$

$$0 \le \sum_{i,j} y_i^{Thr} \le 4 \quad \forall i \qquad 0 \le \sum_{i,j} y_i^{Trp} \le 2 \quad \forall i$$

$$0 \le \sum_{i,j} y_i^{Tyr} \le 4 \quad \forall i \qquad 0 \le \sum_{i,j} y_i^{Val} \le 6 \quad \forall i$$

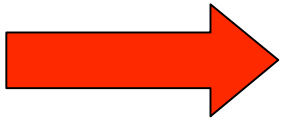
- No. of Cys fixed at 6 (3 S-S bridges)
- No. of Pro (inflexible) fixed at 5 (same as native sequence)
- Max. Trp set to 2 to allow greater flexibility
- No constraint on Gly

Amino acid occurrence of the 97 hβD homologs from PSI-BLAST

# De Novo Design of hβD-2

## In Silico Sequence Selection

- Pos 1, 3, 12, 31, and 34 fixed at Gly
- Pos 5, 17, 21, 33, and 41 fixed at Pro
- Pos 8, 15, 20, 30, 37, and 38 fixed at Cys
- Full combinatorial optimization (all 20 amino acids allowed) for other positions



Complexity: $20^{25}$ or $3.4\times10^{32}$ sequences

Global energy minimum solution:

| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Gly | Arg | Gly | Tyr | Pro | Arg | Asn | Cys | Asp | Thr | Lys | Gly | Tyr | Tyr | Cys | Tyr | Pro | Met | Ala | Cys |

| 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 | 31 | 32 | 33 | 34 | 35 | 36 | 37 | 38 | 39 | 40 | 41 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Pro | Arg | His | Arg | His | Phe | Phe | His | Met | Cys | Gly | Met | Pro | Gly | Phe | Phe | Cys | Cys | Ala | His | Pro |

# 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 = $1.0 \times 10^{13}$

- 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 = $3.3 \times 10^{19}$

- 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 = $1.7 \times 10^{31}$

- 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 = $3.4 \times 10^{45}$

# Quadratic Assignment Like Formulation Comparison

## Computation time comparison of the 12 formulations:

| | Sequence Search space | Formulations | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | **F1** | **F2** | **F3** | **F4** | **F5** | **F6** | **F7** |
| Problem 1 | $1.3x10^{8}$ | 0.14 | 0.30 | 0.05 | 0.04 | 0.05 | 0.15 | 0.23, 0.21 |
| Problem 2 | $1.0x10^{13}$ | 1.93 | 34874 | 12.80 | 65.04 | 13.23 | 2.16 | 44.02, 3.01 |
| Problem 3 | $3.3x10^{19}$ | 3.01 | 70.14% gap | 137.85 | 2052.2 | 278.0 | 3.22 | 64.39, 2.87 |
| Problem 4 | $1.7x10^{31}$ | 38.14 | - | - | - | - | 31.67 | - , 29.06 |
| Problem 5 | $3.4x10^{45}$ | 74713 | - | - | - | - | 30006 | - , 65575 |

No cutoff for triangle inequalities

Cutoff for triangle inequalities = -40

Obtained after 100,000 CPU sec

| | Sequence search space | Formulations | | | | |
|---|---|---|---|---|---|---|
| | | **F8** | **F9** cutoff=-40 | **F10** cutoff =-40 | **F11** | **F12** cutoff = -40 |
| Problem 1 | $1.3x10^{8}$ | 0.16 | 0.11 | 0.16 | 0.17 | 0.11 |
| Problem 2 | $1.0x10^{13}$ | 2.15 | 2.26 | 2.01 | 2.52 | 2.10 |
| Problem 3 | $3.3x10^{19}$ | 2.94 | 3.31 | 3.03 | 3.43 | 3.04 |
| Problem 4 | $1.7x10^{31}$ | 31.08 | 35.48 | 35.92 | 25.00 | 36.15 |
| Problem 5 | $3.4x10^{45}$ | 32657 | 52276 | 61872 | 24388 | 57569 |

67% reduction in CPU time

F1: Base case - original $O(n^2)$ formulation from Klepeis et al. (2003)(2004)

F2: original $O(n^2)$ formulation without RLTs

F3: O(n) formulation from Oral and Kettani (1990)(1992)

F4: O(n) formulation from Oral and Kettani (1990)(1992)

F5: O(n) formulation from Pardalos et al. (2004)

F6: original $O(n^2)$ formulation with inequality RLT constraints

F7: original $O(n^2)$ formulation with inequality RLT constrains and triangle inequalities

F8: original $O(n^2)$ formulation with inequality RLT constraints and preprocessing

F9: original $O(n^2)$ formulation with inequality RLT constraints and triangle inequalities and preprocessing

F10: original $O(n^2)$ formulation with triangle inequalities

F11: original $O(n^2)$ formulation with preprocessing

F12: original $O(n^2)$ formulation with triangle inequalities and preprocessing

# Sequence selection: Comparison

- Test case: human $\beta$ defensin-2 (41 Amino acids; 3 C-C)

First problem

- mutate all positions except CYS. Allow all 20 amino acids for each mutated position
- no biological constraints

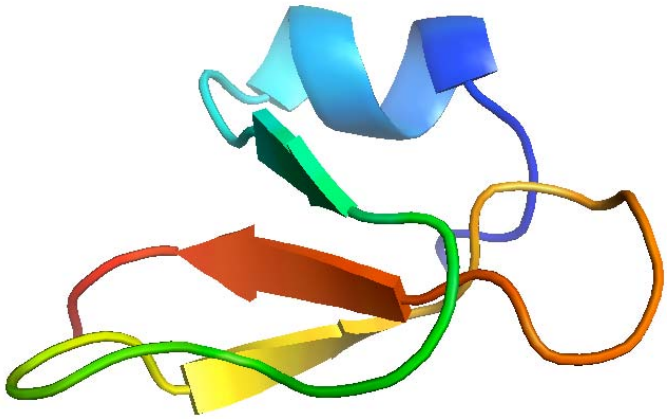
| First Problem | | | |
|---|---|---|---|
| Problem complexity | No. of linear biological constraints | CPU times[s] | |
| | | old formulation | new formulation |
| **$3.4 \times 10^{45}$** | none | **53,263** | **649** |

Second problem

- mutation set derived from SASA patterning

SASA < 20%: core. Allow only hydrophobic amino acids.
SASA 20-50%: intermediate. Allow all amino acids except CYS
SASA >50%: surface. Allow only hydrophilic amino acids

- 49 biological constraints



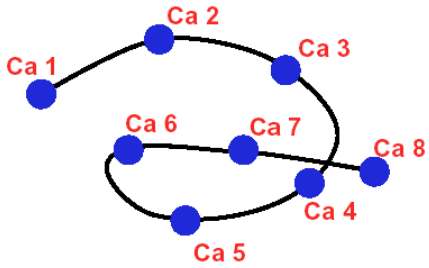
Bounds on charges, hydrophobic content, and amino acid occurrence from PSI-BLAST

| Second Problem | | | |
|---|---|---|---|
| Problem complexity | No. of linear biological constraints | CPU times[s] | |
| | | old formulation | new formulation |
| **$6.4 \times 10^{37}$** | 49 | **4,578** | **14** |

-CPU times generated using CPLEX 9.0 on one single Pentium IV 3.2 GHz processor

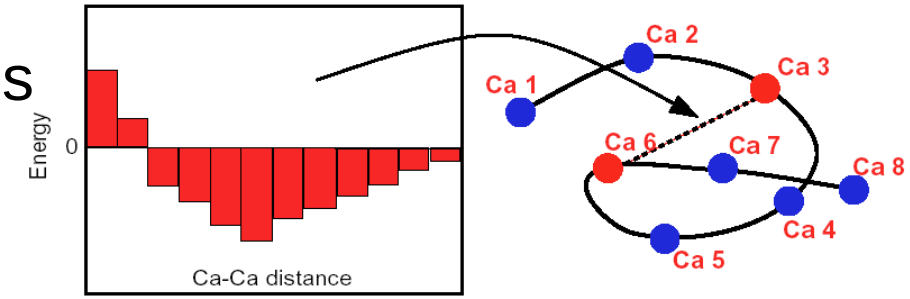
# De Novo Protein Design Framework

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




## Sequence selection

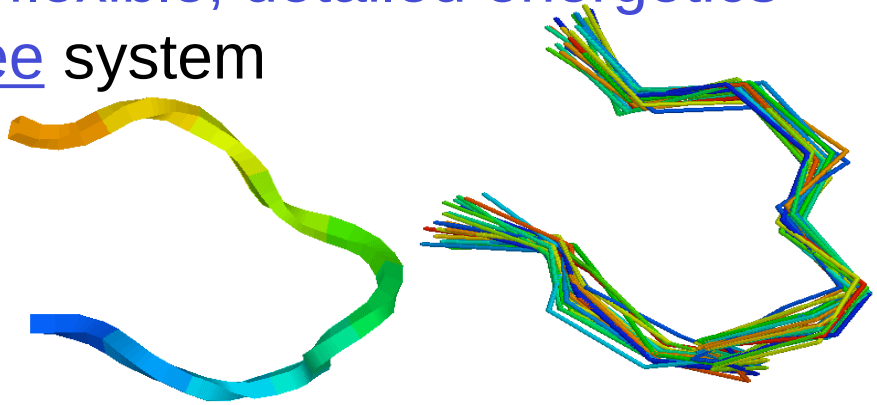
- Identify target template for desired fold; specify coordinates of backbone
- Identify possible residue mutations
- Introduce distance dependent pairwise potential based on Ca
- Generate rank-ordered energetic list from mixed-integer linear (MILP)




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



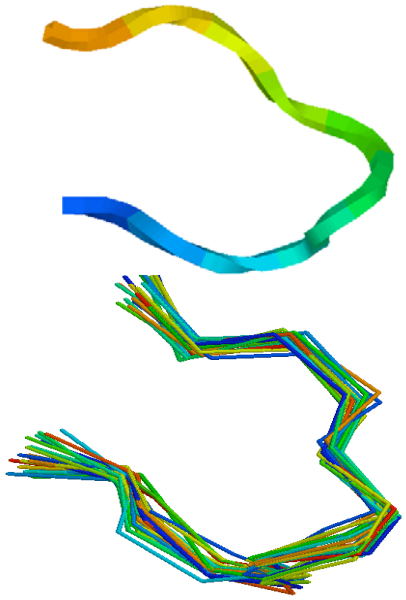
# Fold Validation : Astro-Fold based

## 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

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

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

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

# Enhanced ASTRO-FOLD

**Helix Prediction**
-Detailed atomistic modeling
-Simulations of local interactions
(Free Energy Calculations)

**β-sheet Prediction**
-Novel hydrophobic modeling
-Predict list of optimal topologies
(Combinatorial Optimization)

α proteins

α/β proteins

**Interhelical Contacts**
-Maximize common residue pairs
-Rank-order list of topologies
(MILP Optimization Model)

**Flexible Stems Loop Prediction**
-Dihedral angle sampling
-Discard conformers by clustering
(Novel Clustering Methodology)

**Derivation of Restraints**
-Dihedral angle restrictions
-Cα–Cα distance constraints

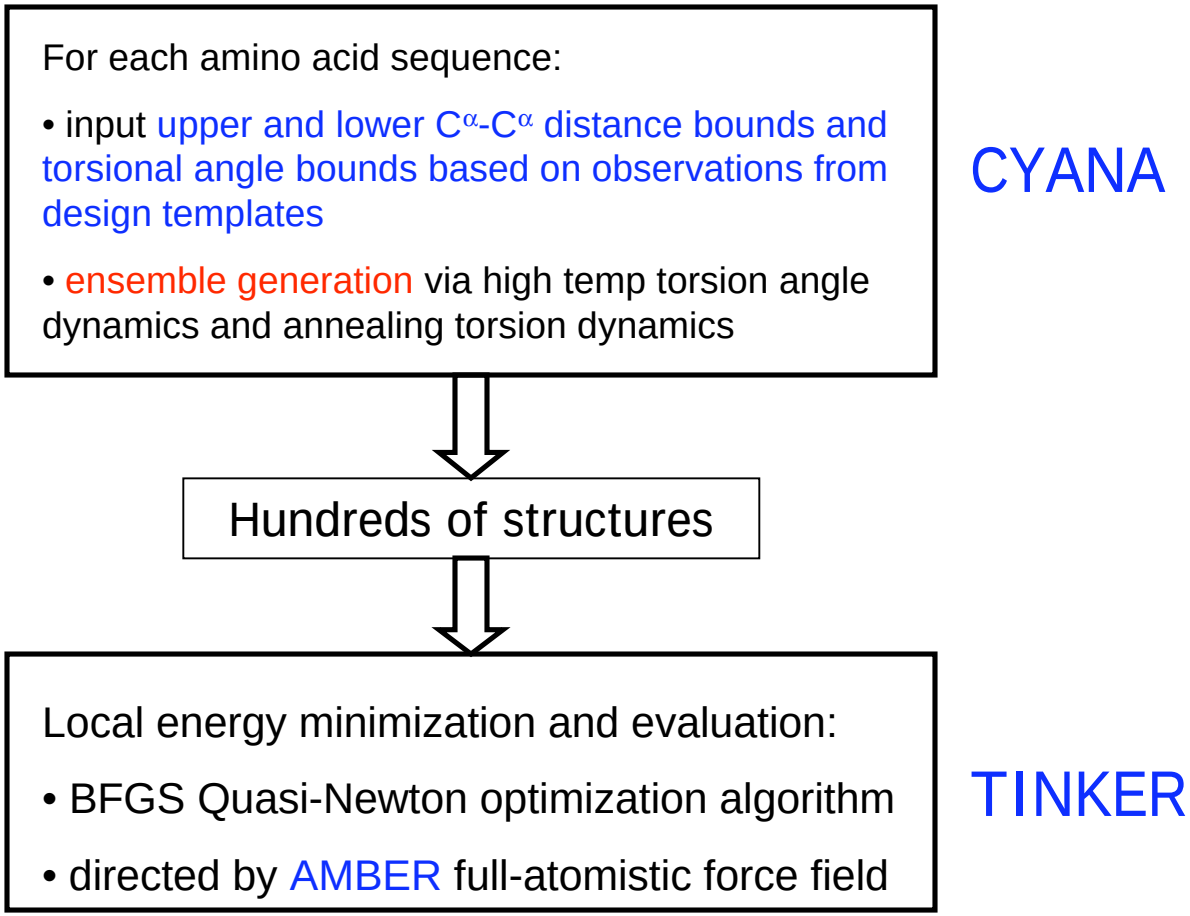
**Improved Distance Restraints**
-Iterative LP-based bound tightening approach

**Tertiary Structure Prediction**
-Structural data from previous stages
-Prediction via novel solution approach
(Global Optimization and Torsional Angle Dynamics)

**Force Field for High and Medium Resolution Decoys**
-Novel linear programming approach
-Distinguishes high resolution structures
(Large-scale linear programming)

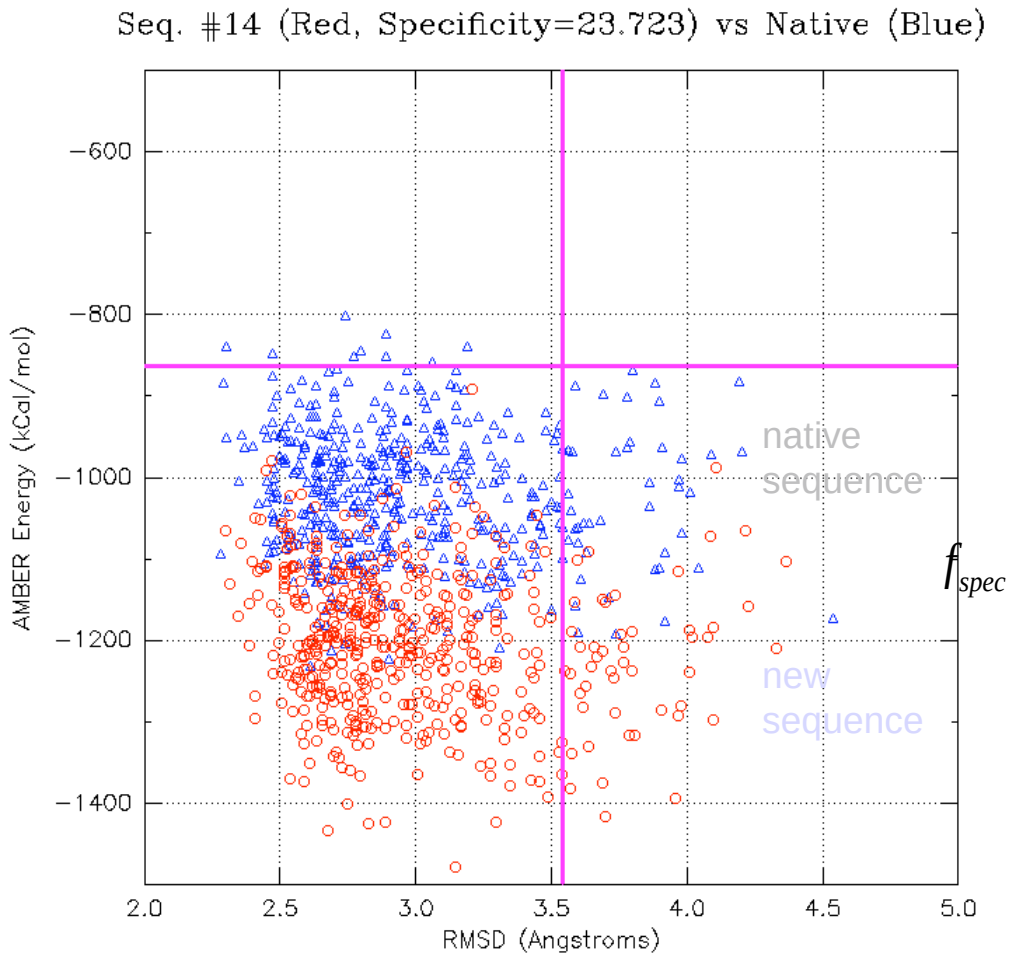
# Fold Validation: NMR like framework

- 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



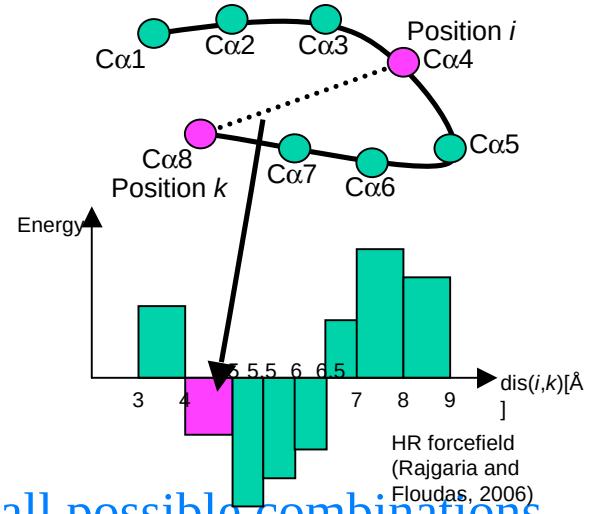

$$f_{spec} = \frac{\sum_{i \in \text{conformers of new sequence}} \exp(-\frac{E_i}{kT})}{\sum_{i \in \text{conformers of native sequence}} \exp(-\frac{E_i}{kT})}$$

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

# De Novo Design of Inhibitors for Complement 3: Compstatin variants

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

Prof. D. Morikis (UC, Riverside)

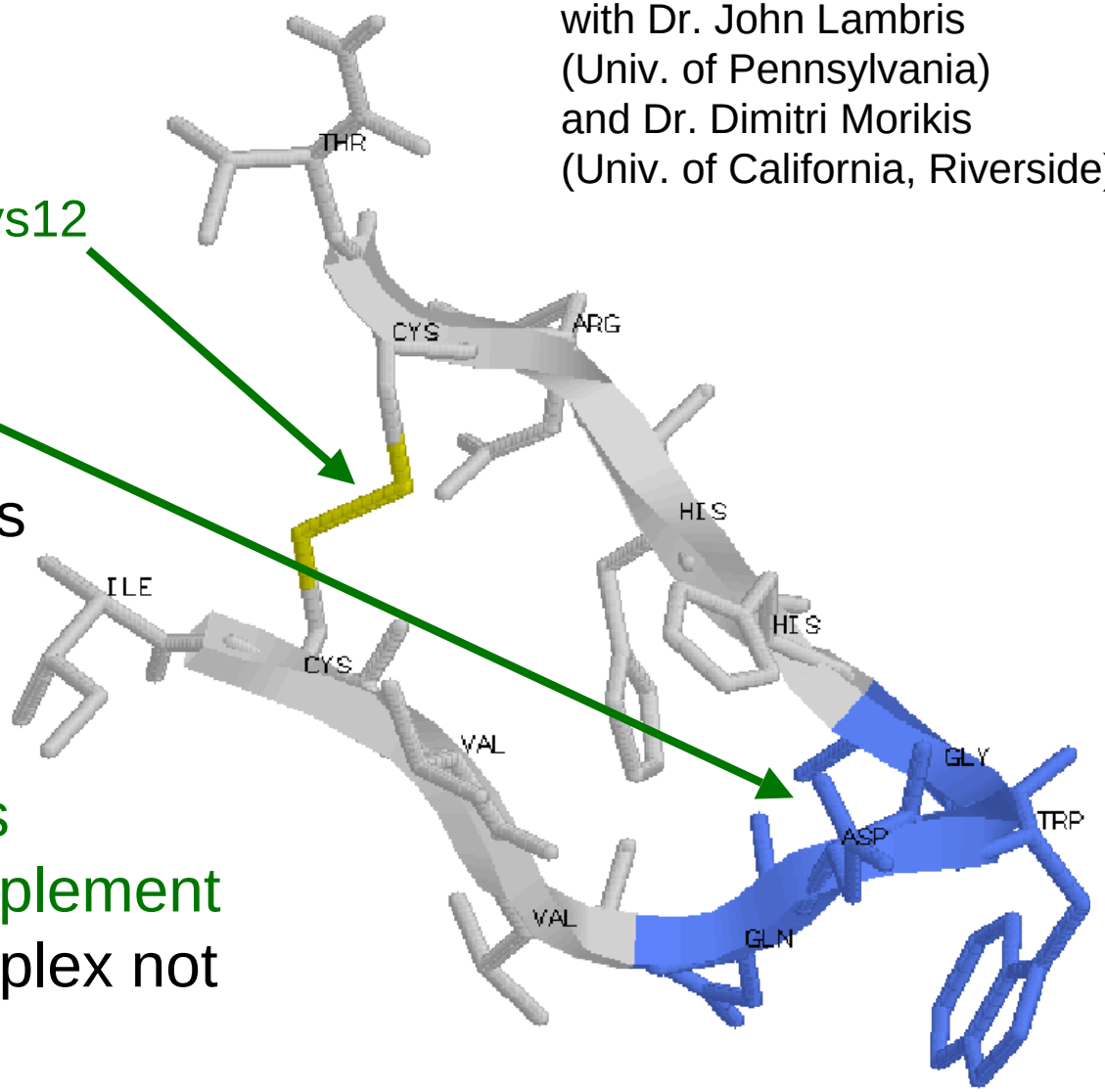
# Compstatin

## 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)

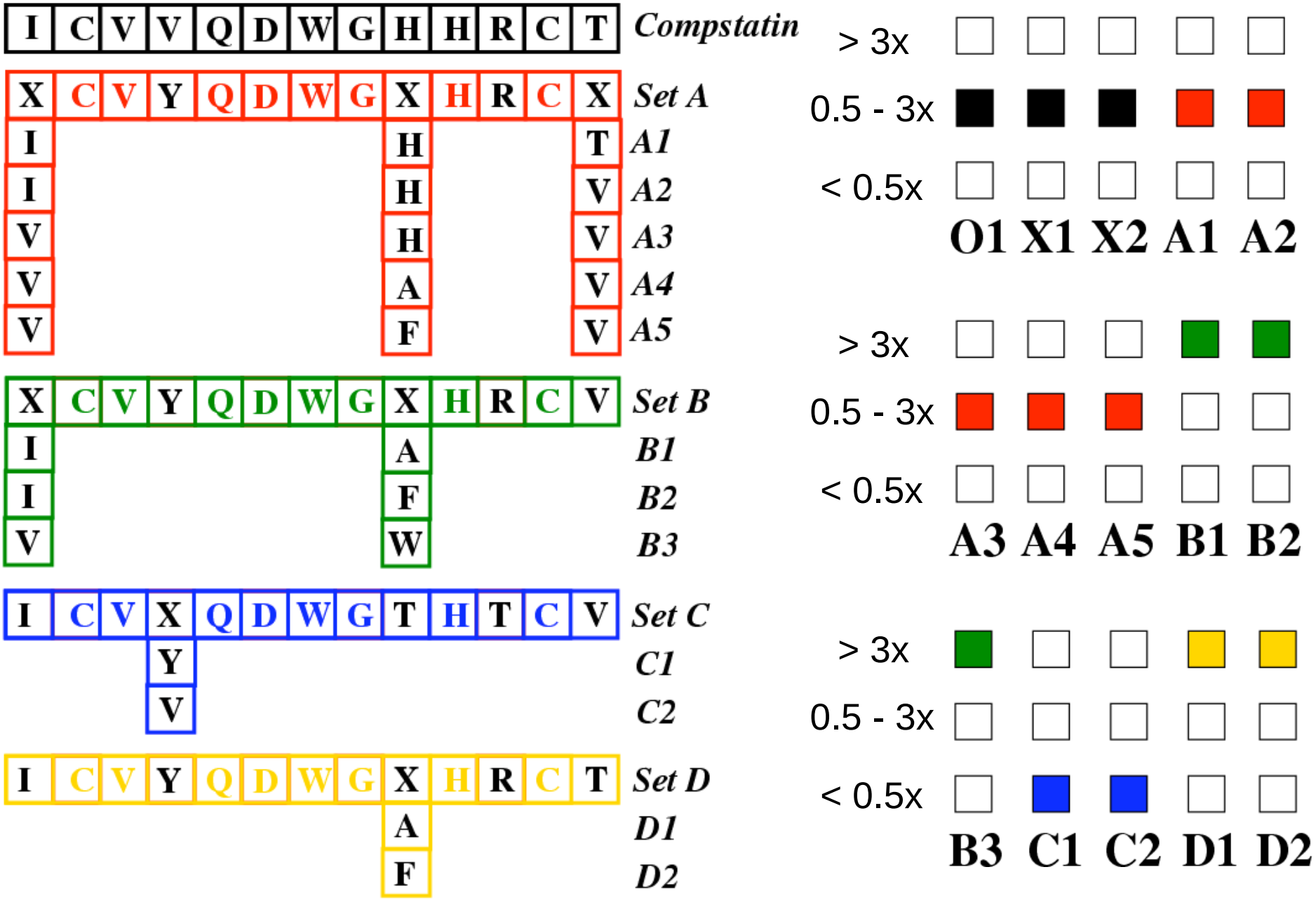
Consensus results from top sequences

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

Key finding from computations

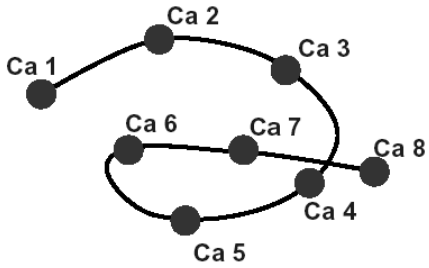
- His conserved at position 10
- Position 11 provides most variation : maintain Arg
- Selections at positions 4 and 9 allow for turn flexibility

# Compstatin Analogs



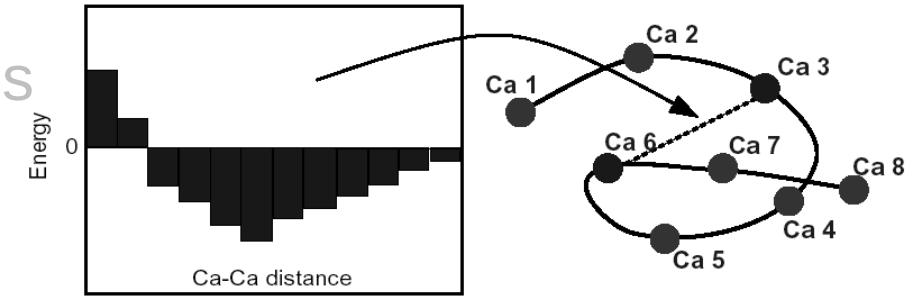
# De Novo Protein Design Framework

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




## Sequence selection

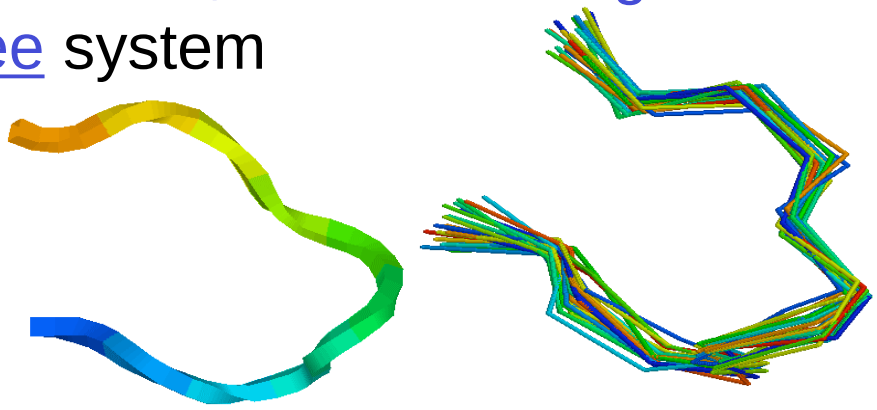
- Identify target template for desired fold; specify coordinates of backbone
- Identify possible residue mutations
- Introduce distance dependent pairwise potential based on Ca
- Generate rank-ordered energetic list from mixed-integer linear (MILP)




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

## Fold Validation

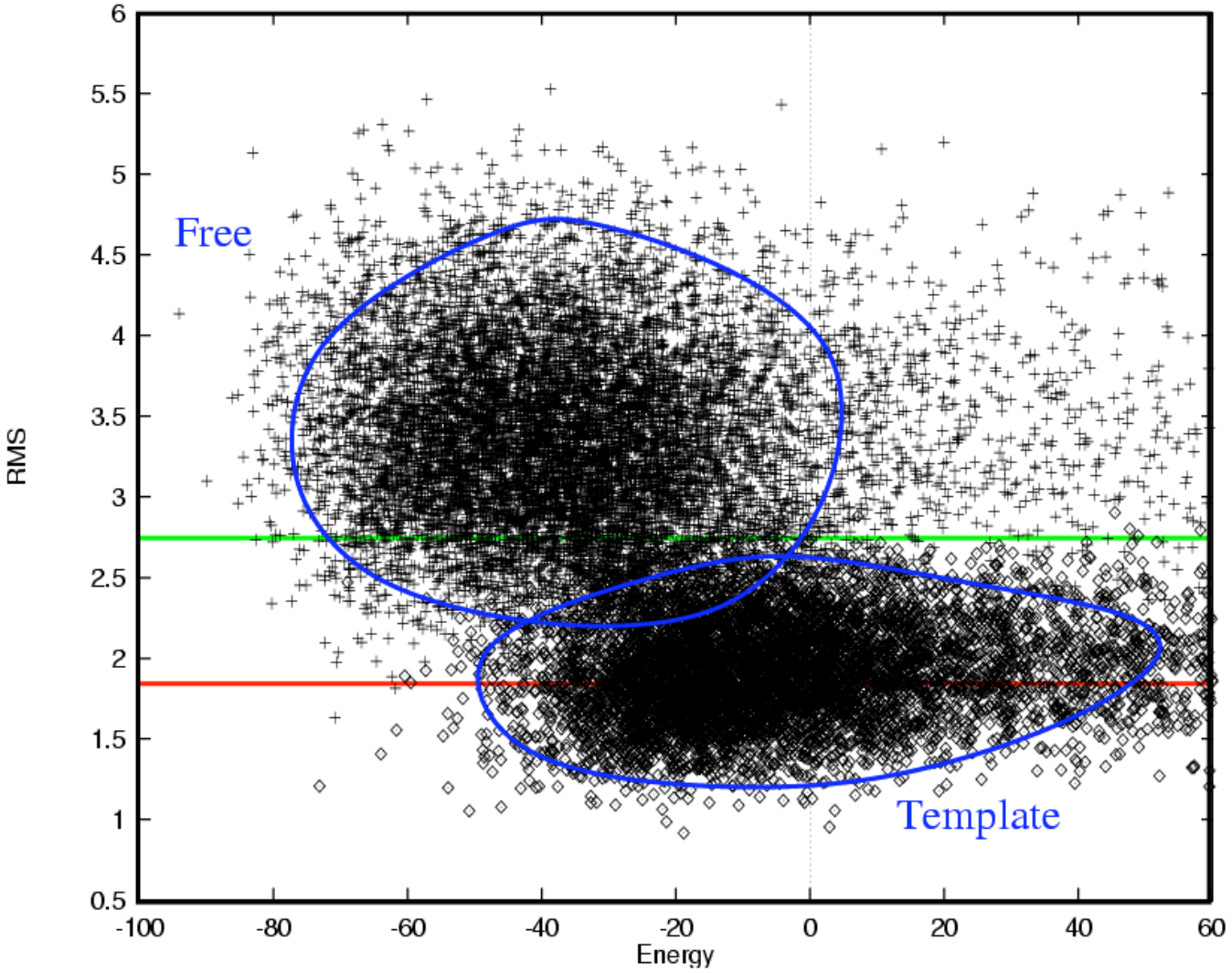
- Model selected sequences using flexible, detailed energetics
- Employ global optimization for free system
- Employ global optimization for system constrained to template
- Calculate relative probability for structures similar to desired fold



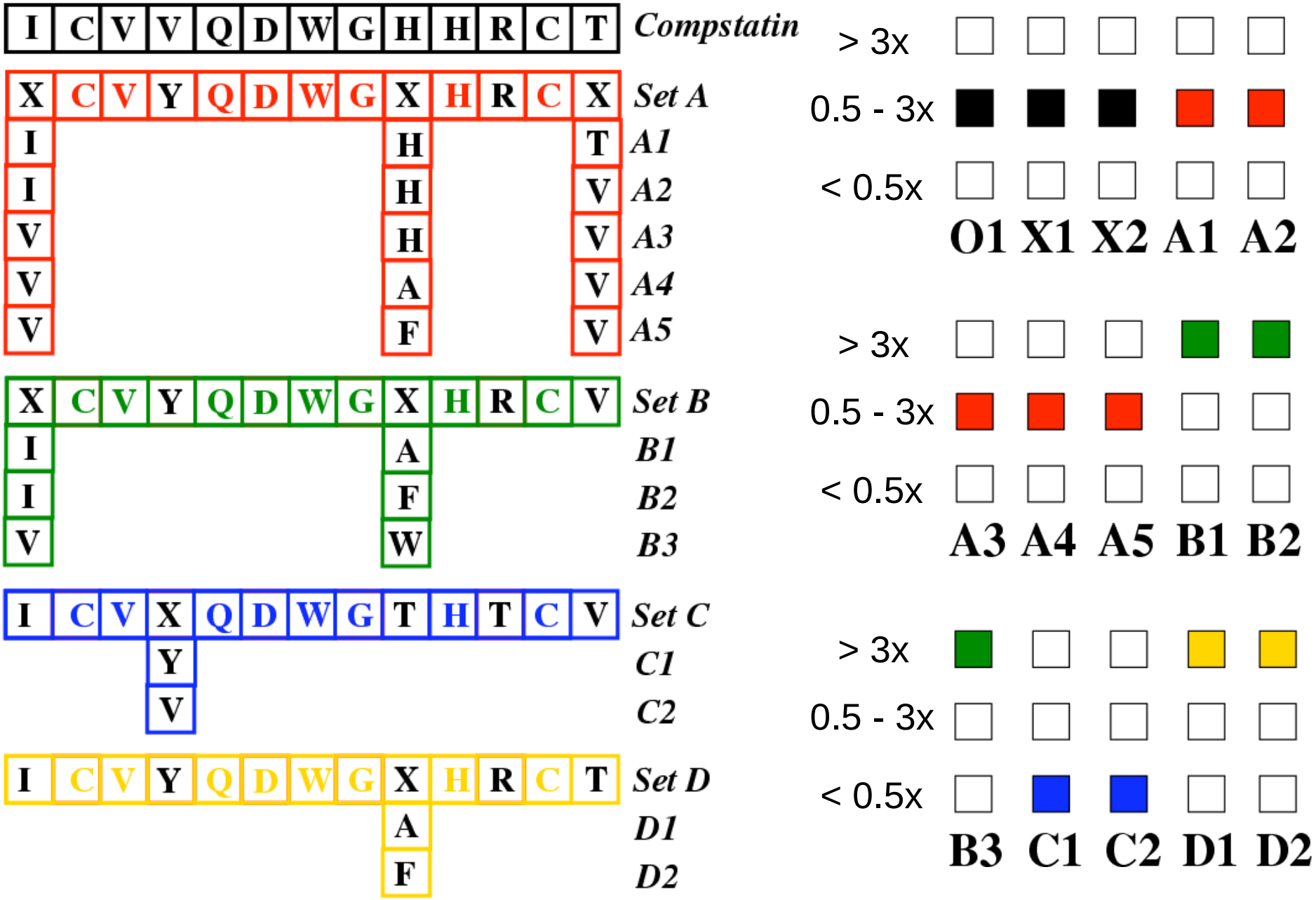
# Fold Specificity : Compstatin

Determine ensemble for Free & Template systems

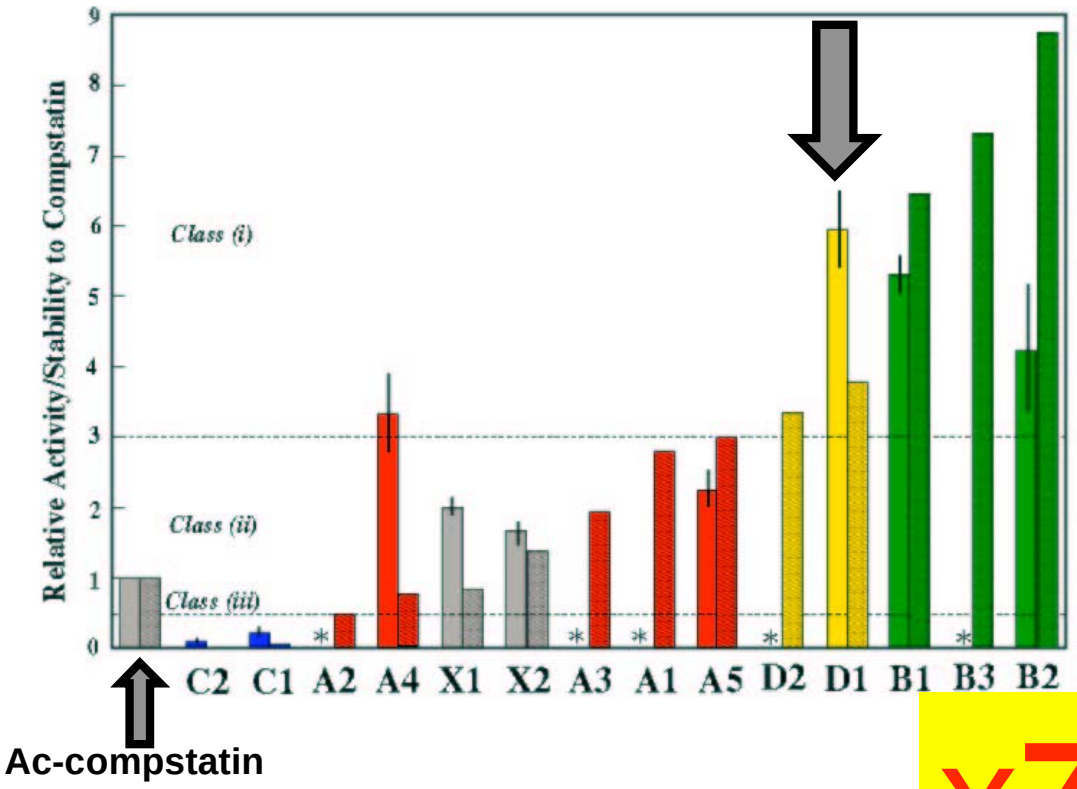
Find probability for portion of Free ensemble within some deviation of Template ensemble



# 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)

# Redesign of Complement 3a

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

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

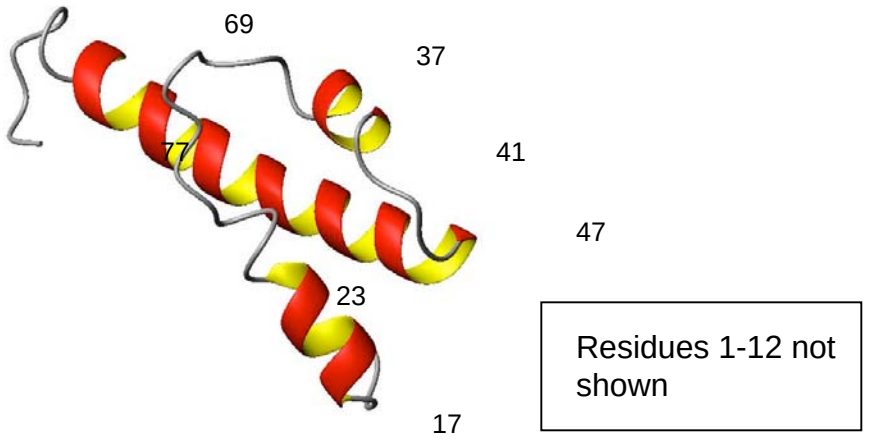
  - extensive sequence-activity studies by Ember *et al*. (1991)

  - 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**

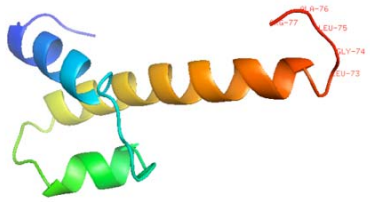
  - goal: **identify peptides that are more active than natural C3a**



# De novo design of C3a

- We used 3 sets of design templates:

1. Single structure from X-ray crystallography

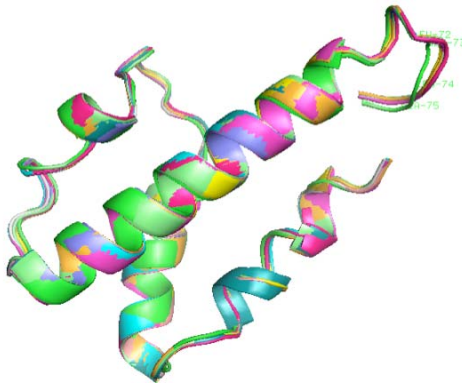


- Huber *et al*., 1980
- Resolution: 3.2Å
- Residue 1 to 12 missing
- Side-chain information is also missing

2. Flexible templates from MD simulations with GB implicit solvation

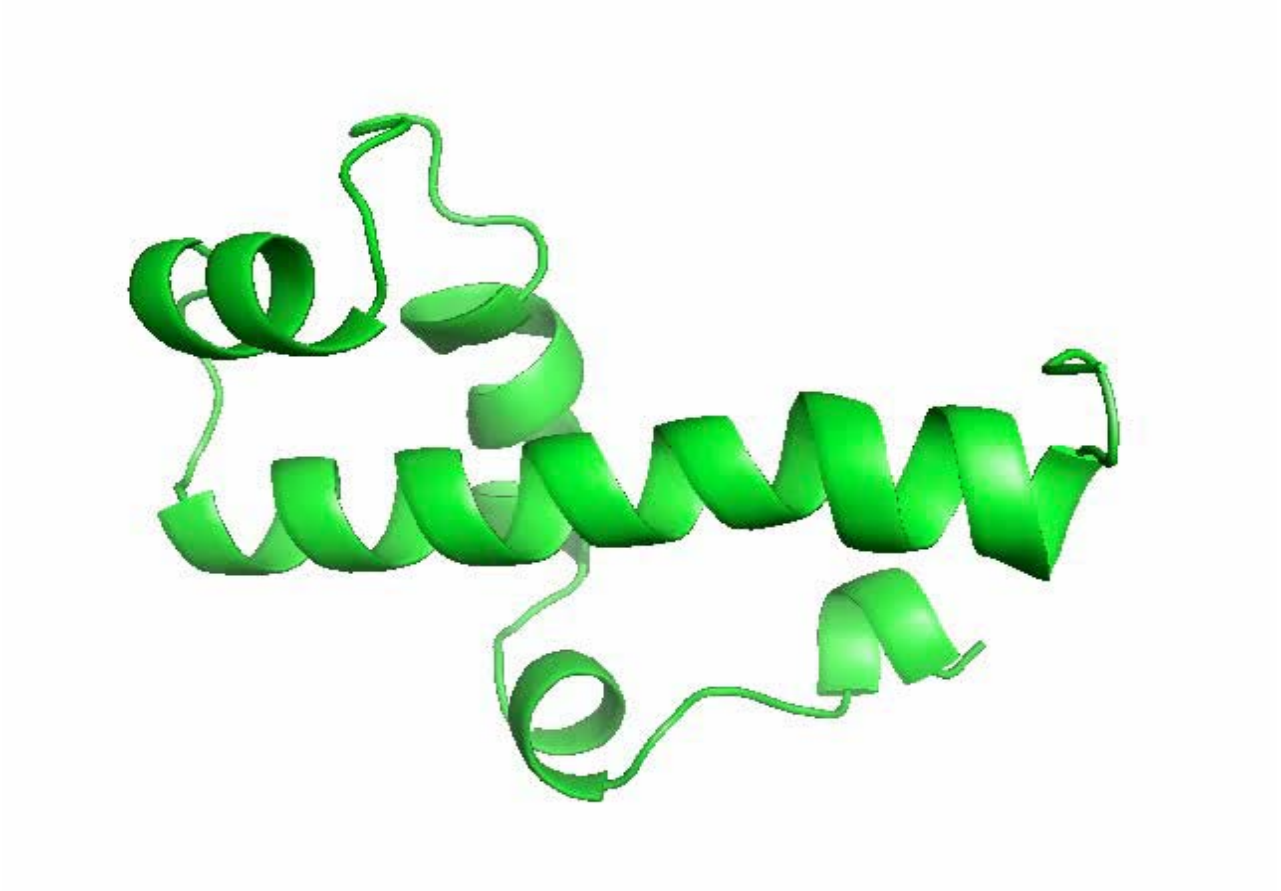
- Initial structure: composite of C3a domain of C3's crystal structure (Janssen *et al*., 2005) for $Val^{1}$-$Ala^{70}$ and Huber *et al*.'s crystal structure for $Ser^{71}$-$Arg^{77}$
- **Starting from 10ns, one structure generated at each 1ns increment.**
- **11 flexible template structures in total**

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**

# Flexible design template with multiple structures for C3a

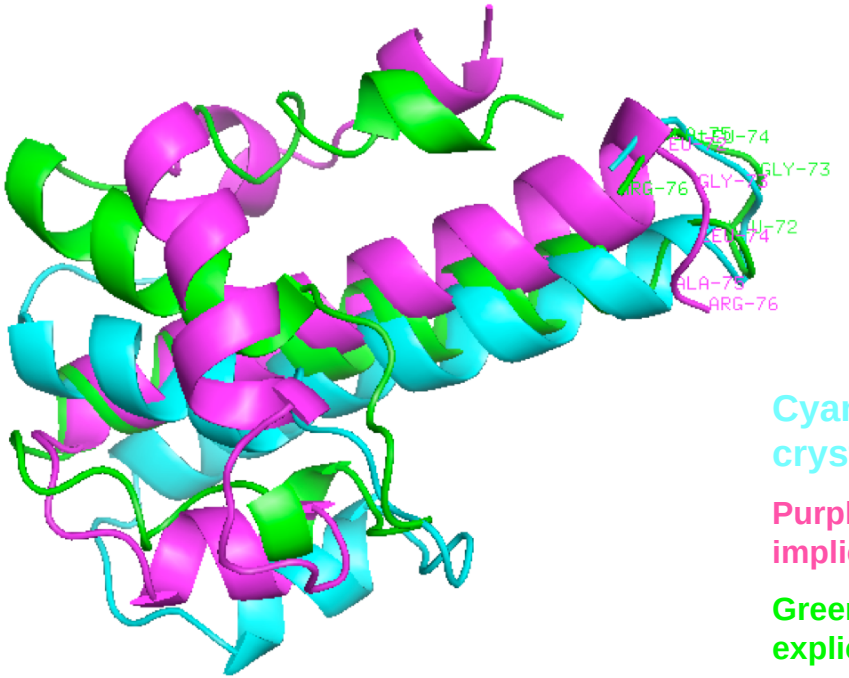


Green: from MD simulations with GB implicit solvation

Magenta: from MD simulations with explicit water molecules

# De novo design of C3a

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




**Cyan: single structure from X-ray crystallography**

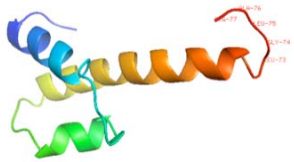
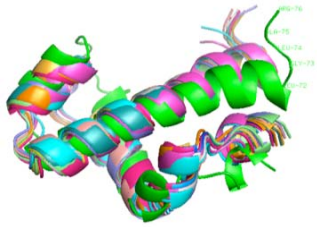
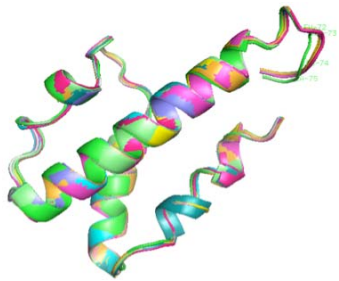
**Purple: structure at 5 ns from MD simulations with GB implicit solvation**

**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 $C^{\alpha}$-$C^{\alpha}$ forcefield only | • Basic model for single structure |
| MD simulations with GB implicit solvent  | • HR $C^{\alpha}$-$C^{\alpha}$ forcefield<br>• HR centroid-centroid forcefield | • Weighted average model for multiple structures<br>• Binary distance bin model for multiple structures |
| MD simulations with explicit water  | • HR $C^{\alpha}$-$C^{\alpha}$ forcefield<br>• HR centroid-centroid forcefield | • Weighted average model for multiple structures<br>• 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 |

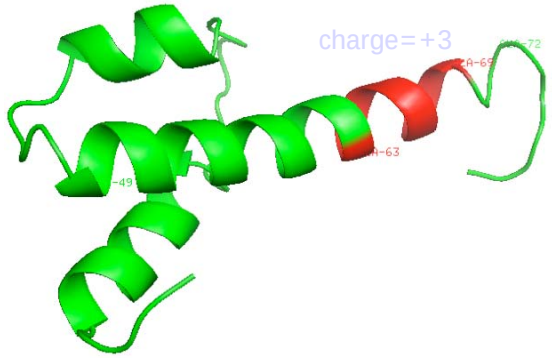
Problem complexity

$=2.59\times10^{9}$

- **Biological constraint**

  - Maintain native charge on helix

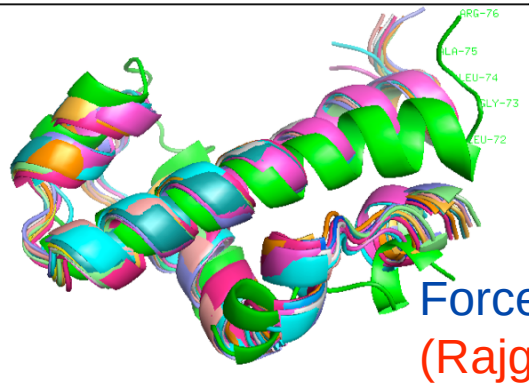
$$\sum_i y_i^{Arg} + \sum_i y_i^{Lys} - \sum_i y_i^{Asp} - \sum_i y_i^{Glu} = 3 \;\; \forall\, 63 \le i \le 69$$



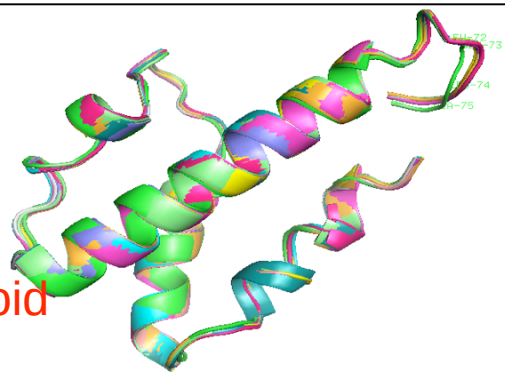

- Fold specificity stage: upper and lower bounds on angles and distances are based on observations about the flexible template(s).

# 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**

# Conclusions

## De Novo Peptide Design : Structure to Function

- Novel method for sequence selection
  - Distance dependent pairwise interaction energy
  - MILP reformulation: Quadratic Assignment-Like
  - RLT constraints
  - Preprocessing via DEE
  - New Formulation
- Quantification of fold specificity
  - Template flexibility
  - Constrained and free energy calculations
  - Ranking of sequence-structure specificity
- Sequence Selection for Compstatin, Human beta defensin-2, C3a, and HIV-1
- Fold specificity for Compstatin analogs, C3a

**Functionally enhanced peptides for C3 inhibition**

# Discovery in Proteomics: De Novo and Hybrid Methods via Tandem Mass Spectrometry


ChE
PrincetonUniversity

**Professor Christodoulos A. Floudas**

Department of Chemical Engineering

Program in Applied and Computational Mathematics

Department of Operations Research and Financial Engineering

Center for Quantitative Biology

**Princeton University**

# Peptide Identification In Proteomics

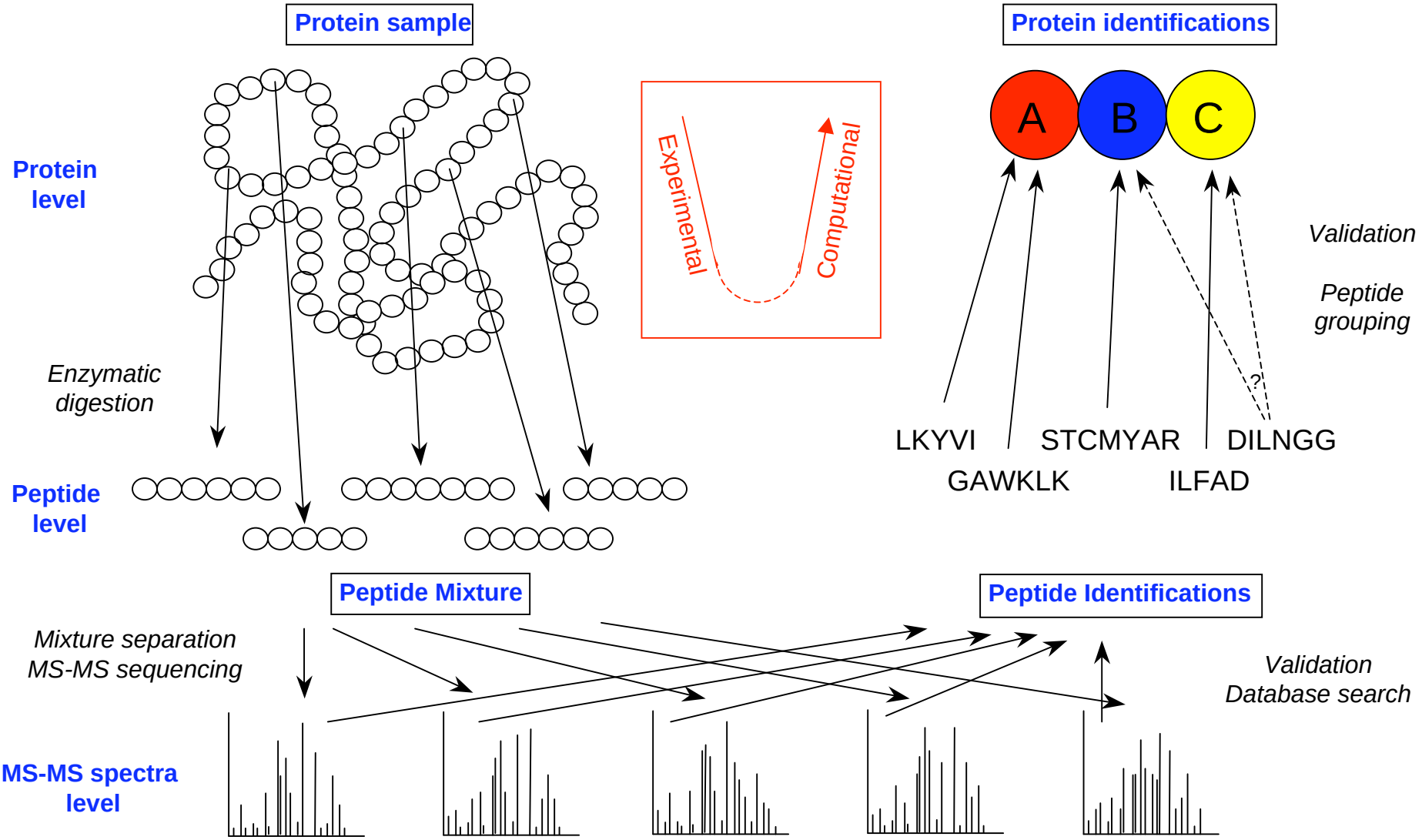
## Relevant Publications


- P.A. DiMaggio and C.A. Floudas, A mixed-integer optimization framework for de novo peptide identification, *AIChE Journal*, 53(1), 160-173 (2007).
- P.A. DiMaggio and C.A. Floudas, De novo peptide identification via tandem mass spectrometry and integer linear optimization, *Anal. Chem.*, 79, 1433-1446 (2007).
- P.A. DiMaggio, C.A. Floudas, B. Lu, and J.R Yates, A hybrid methodology for peptide identification using integer linear optimization, local database search, and QTOF or OrbiTrap tandem mass spectrometry, *J. Proteome Res.*, 7, 1584-1593 (2008).

# Presentation Outline

- Introduction to **proteomics** and review of **peptide and protein identification** using **tandem mass spectrometry**
- Survey of existing **de novo** and **database** algorithms for peptide identification
- **De Novo** approach for **peptide identification**, **PILOT**
- Hybrid approach based on our **mixed-integer linear optimization** model and algorithmic framework (denoted as **PILOT_SEQUEL**) for **peptide identification**
- **Comparative studies** of **PILOT**, and **PILOT_SEQUEL**, and existing state-of-the-art *database* and *hybrid* methods on tandem MS from **Ion Trap**, **QTOF**, and **OrbiTrap** mass analyzers

# Proteomics: Bottom-Up Peptide and Protein Identification via Tandem MS



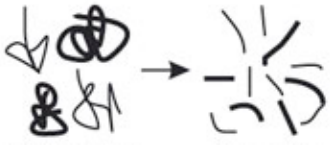
# Problem Introduction and Definition

- Fundamental problem in proteomics:

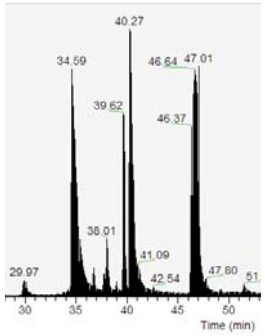
  ***Protein and peptide identification and quantification***

- Advances in **high-throughput** experimentation

  High-performance liquid chromatography (**HPLC**) coupled with tandem mass spectrometry (MS/MS)



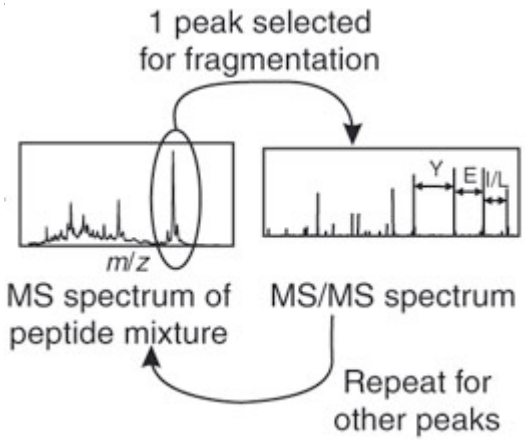
Protein / Peptide Mixture







**MS and MS/MS**

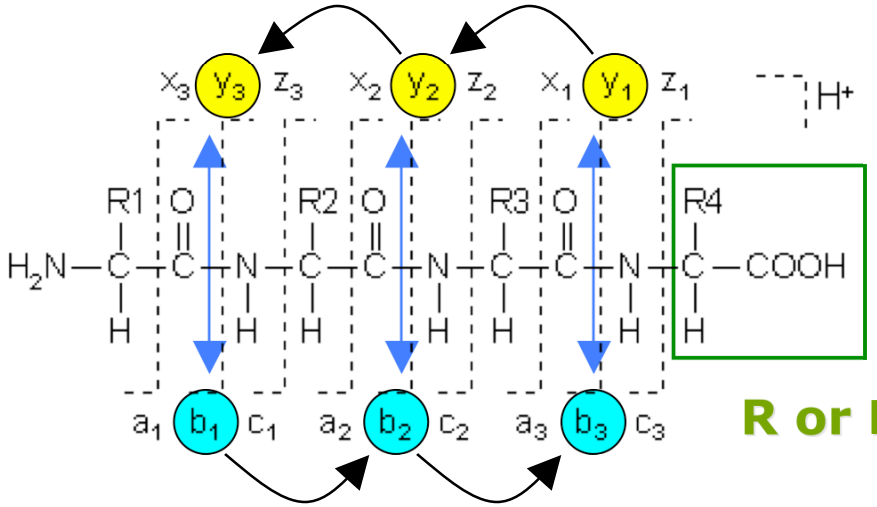



**Electrospray Ionization (ESI)**

***Fragmentation…***

# Tandem MS/MS from CID

❑ **Collision-induced dissociation** (CID) causes a positively-charged peptide to fragment along its backbone and results in several types of **fragment ions** in the tandem mass spectrum (i.e., a, b, c, x, y, z, etc.)




Hypothetical parent peptide*

❑ **Objective**:

*Use these fragment ions to construct the amino acid sequence of the parent peptide*

❑ **Issues**: The **type** of an ion peak (a, b, c, x, y, or z) in a tandem MS is *not known* a priori and the primary sequence of candidate peptide must be derived using **ions of the same type**



*Adapted from http://www.matrixscience.com/help/fragmentation_help.html

# Proteomics

- **Specific Aim 1:** Investigate and develop a **de novo** computational approach for peptide identification based exclusively on information of the ion peaks in the peptide spectrum
- **Specific Aim 2:** Study and develop a new **hybrid** in silico method which will combine the de novo approach of Specific Aim 1 with database techniques for peptide identification
- **Specific Aim 3:** Incorporate **uncertainty** into the de novo framework to address experimental uncertainty in problem parameters
- **Specific Aim 4:** Study and develop computational methods for **protein identification** given the de novo prediction and/or hybrid prediction of the individual peptides
- **Specific Aim 5:** Research and develop computational methods and experimental protocols for **protein quantification**

# Peptide & Protein Identification via Tandem MS

- **Database-based methods**
  - **Correlate the experimental spectra with spectra of peptides/proteins which exist in the databases**
  - SEQUEST – Eng et al. (1994), Mascot – Perkins et. al (1999), SCOPE – Bafna and Edwards (2001), MS-CONVOLUTION and MS-ALIGNMENT – Pevzner et. al (2001), Poptiam – Hernandez et. al (2003)
- **De Novo Methods**
  - **Predict peptides without sequence databases**
  - **Exhaustive listing; sub-sequencing; graphical**
  - **Graph theory and shortest path algorithms**
  - **Graph theory and dynamic programming**
  - **Bayesian scoring of random peptides**
  - Lutefisk – Taylor and Johnson (1997,2001), SHERENGA – Dancik et. al (1999), PEAKS – Ma et al. (2003), NovoHMM – Fischer et al. (2005), PepNovo – Frank and Pevzner (2005), EigenMS – Bern and Goldberg (2006)

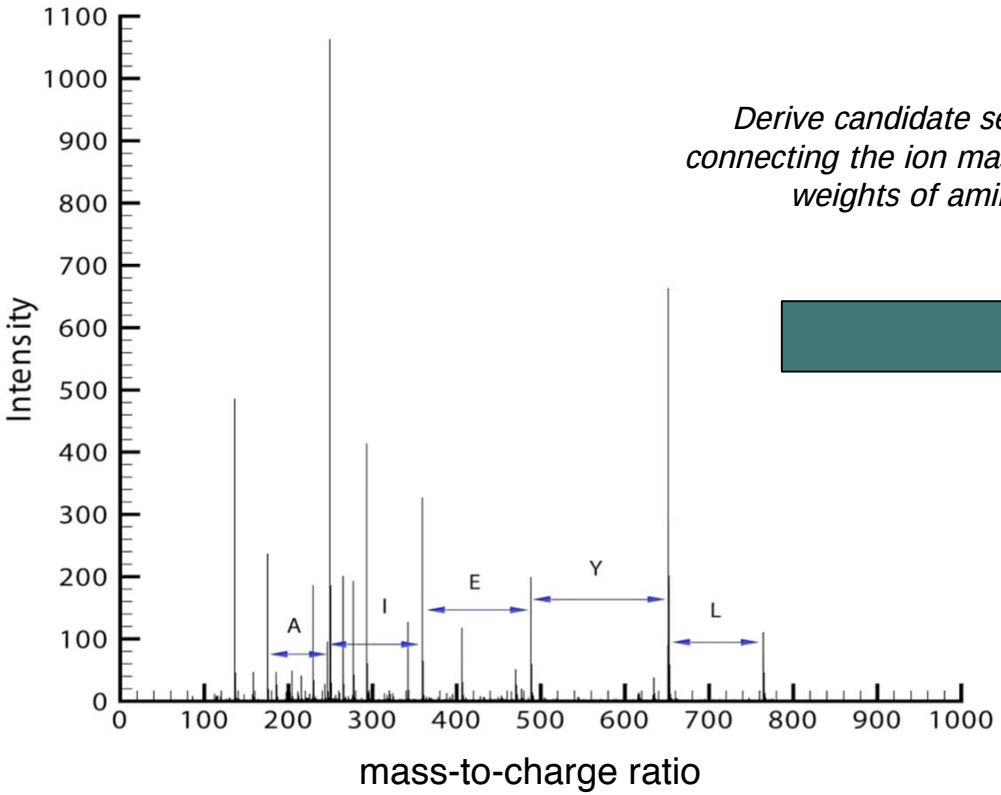
# Challenges

- Tandem MS are **missing ion peaks** due to **incomplete fragmentation** and/or instruments with low mass-to-charge ratio (m/z) cutoff (i.e., ion trap mass analyzers)
- Incorporating parametric **uncertainty** in the measured values for ion peaks during peptide identification
- Existing de novo techniques enumerate an **exhaustive number of candidate sequences** from the tandem mass spectrum
- No straightforward method for including **post-translational modifications** into existing frameworks

# Introduction to De Novo Peptide Identification

## The De Novo Peptide Identification Problem:

Given the **tandem mass spectrum** (MS/MS) of a peptide, derive the primary sequence of the peptide without consulting other sources of information (i.e., protein databases)




*Derive candidate sequences by connecting the ion mass peaks by the weights of amino acids*

**Q**: Which of these possible primary sequences corresponds to the correct peptide?

YLYKNAR

YLFPMTR

YLYELAR

YFEELAR

YEYLLAR

YLYKKGR

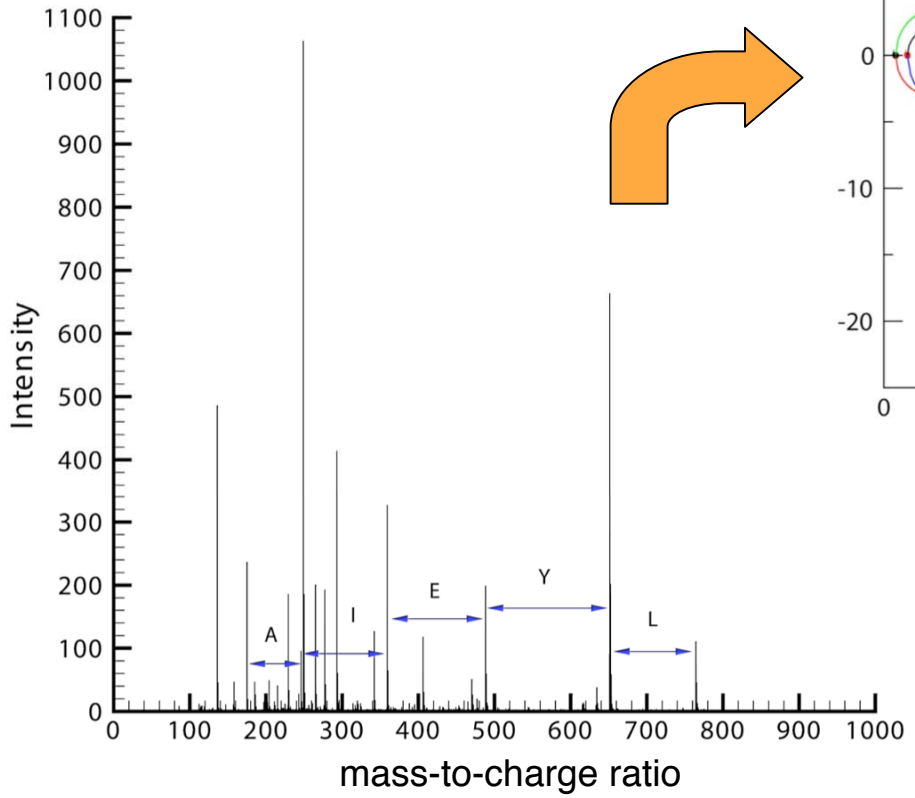
YFEKNAR

YLY[171.06]AAR

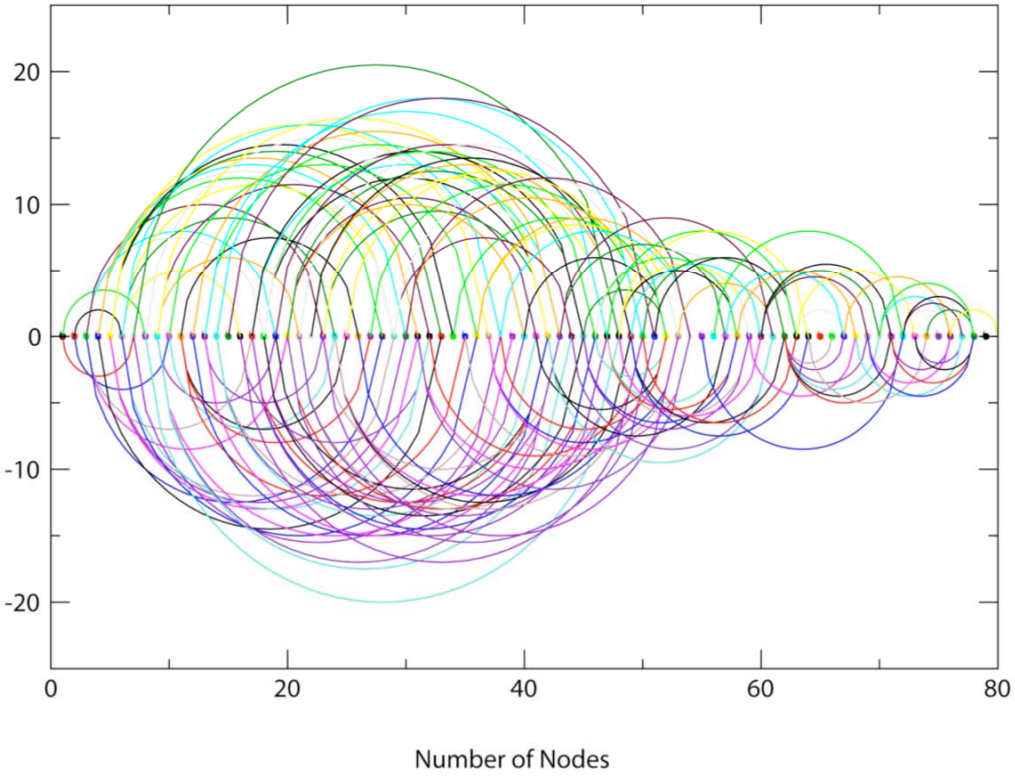
# Traditional De Novo Methods

Transform tandem MS/MS into a spectrum graph, where:

***paths on the graph = amino acid sequences***




Spectrum Graph Approach*



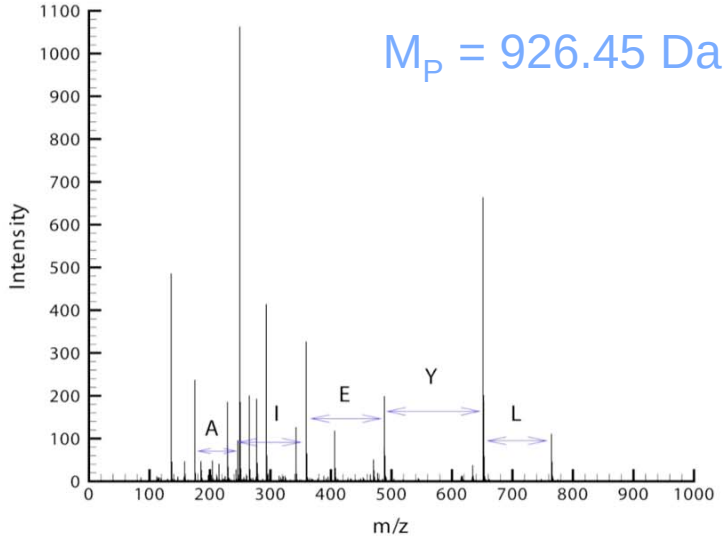

- ❑ Solve via dynamic programming
- ❑ Nodes assigned probabilistic weights
- ❑ Highest scoring path is selected

* Taylor and Johnson (1997,2001), Dancik et. al (1999), Fernandez de Cossio et. al (2000), Chen et. al (2001), Lubeck et. al (2002), Cannon and Jarman (2003), Chen and Bingwen (2003), Jarman et. al (2003), Frank and Pevzner (2005), Bern and Goldberg (2006)

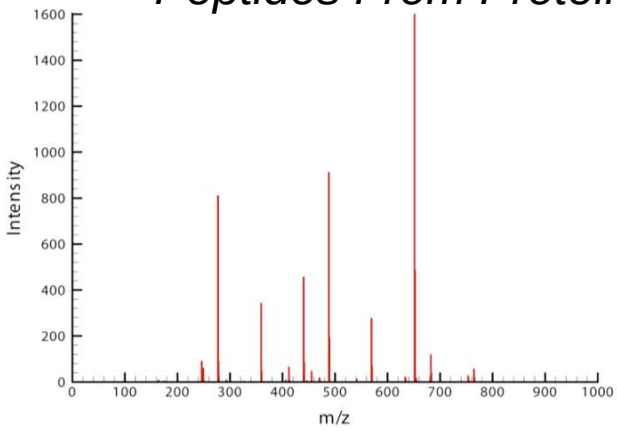
# Database Methods

*Peptides From Protein Database*

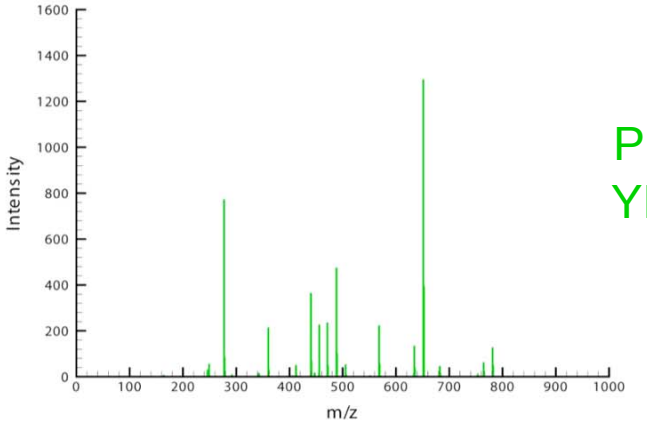
Raw Tandem MS/MS for YLYEIAR



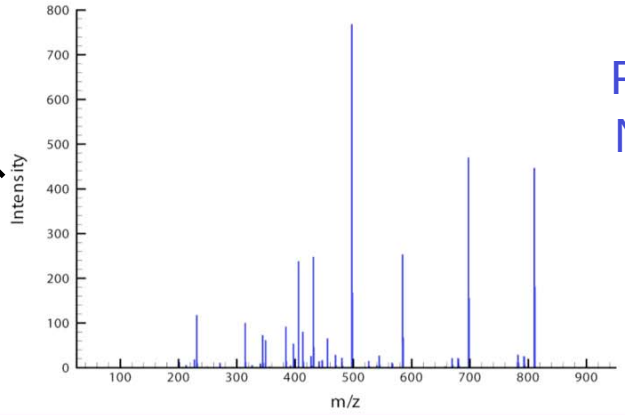

$M_P$ = 926.45 Da




Predicted*
YLYEIAR




Predicted*
YLYQNVK




Predicted*
NRIISLLV

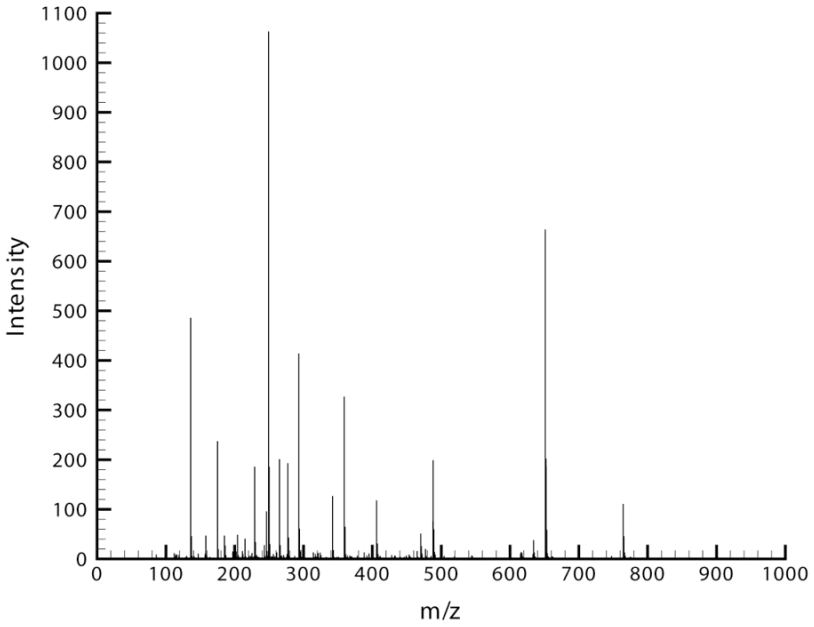
Which ***predicted*** spectrum matches the experimental spectrum under question?

*Predicted spectra generated with MassAnalyzer 1.03

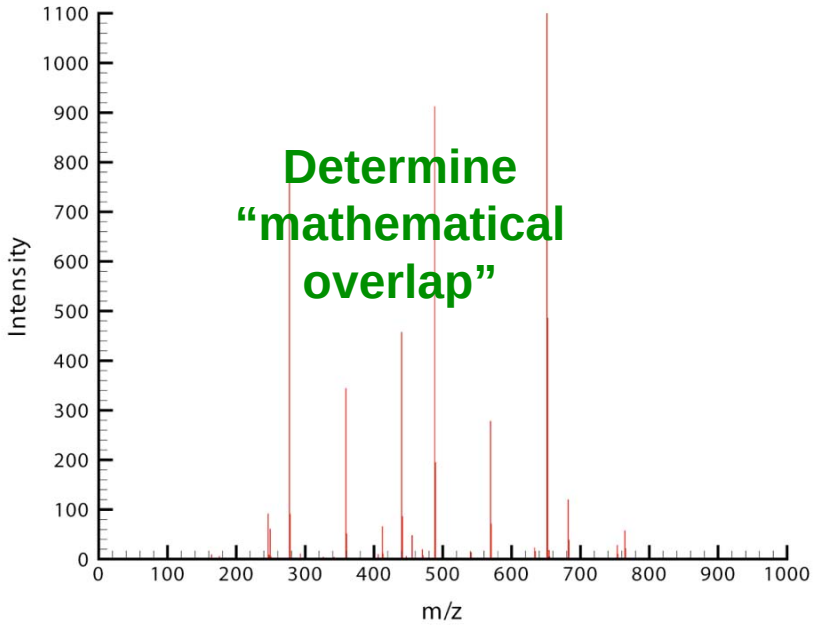
# Database Methods (cont'd)

- **Cross-Correlation** (e.g., SEQUEST*)

Experimental Spectrum → x




Predicted Spectrum → y




**Determine "mathematical overlap"**

$$R_\tau = \sum_{i=0}^{n-1} x[\,i\,]\, y[\,i + \tau\,]$$

Displacement value

$$R_\tau \leftrightarrow X_r Y_r^*$$

Discrete Fourier Transforms

# Database Methods (cont'd)

- **Probabilistic Matching*** (e.g., Mascot, SCOPE)

Predict primarily y- and b-ions, and their **offsets**, based on the following formulae:

$$b_i = \sum_{j=1}^{i} mass(AA_j) + 1 \qquad y_{n-i} = 19 + \sum_{j=i+1}^{n} mass(AA_j)$$

**Q**: Is ion **match** with experimental spectrum → Actual?

**Q**: Is ion **match** with experimental spectrum → Random?

**"A"**:

- **Likelihood ratio hypothesis test** (Bafna and Edwards (2001), Havilio et. al (2003))
- **Null hypothesis** (Sadygov and Yates (2003))
- Integration of spectral dependencies into model (Bafna and Edwards (2001), Havilio et. al (2003))
- Empirically estimated probabilities

*Perkins et. al (1999), Bafna and Edwards (2001), Pevzner et. al (2001), Havilio et. al (2003), Hernandez et. al (2003), Sadygov and Yates (2003)

# Drawbacks of Existing Methods

## De Novo Methods

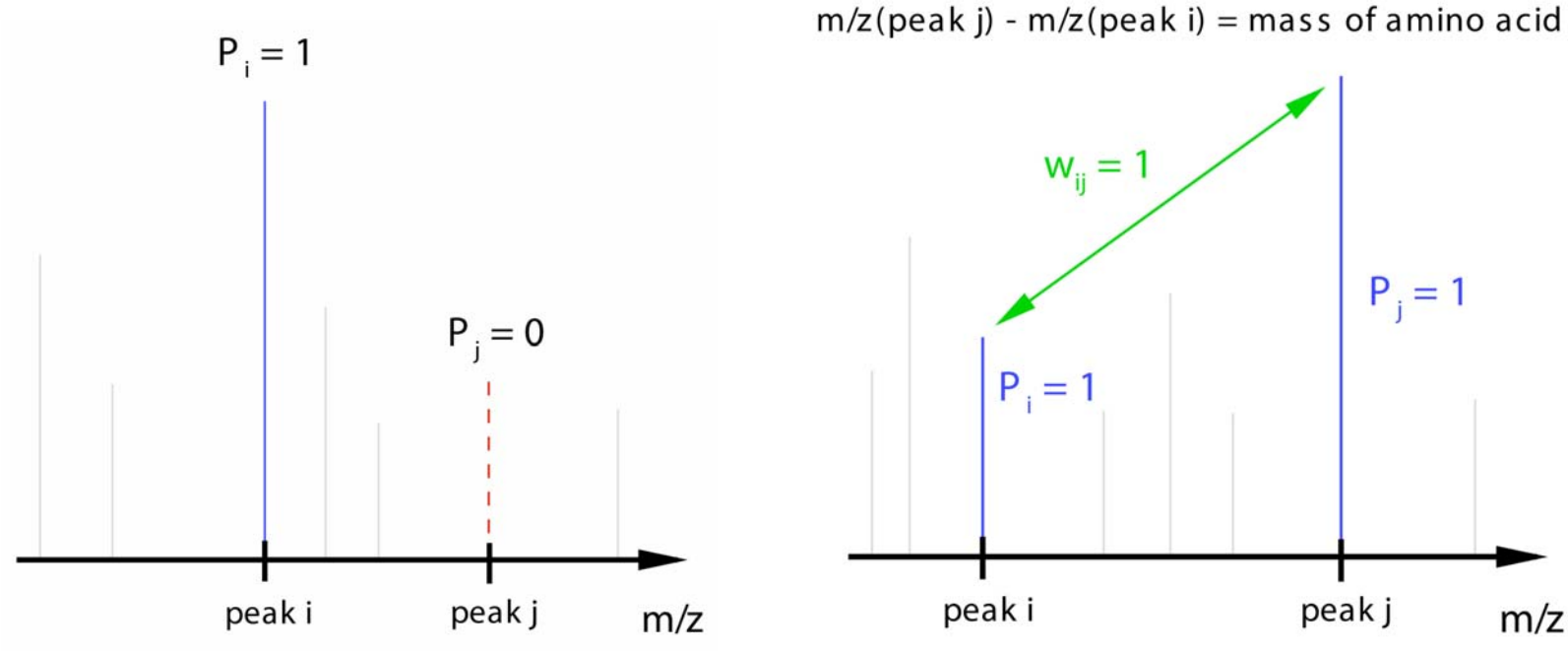
- Exhibit variable prediction **accuracies**
- Computationally **intensive** → exhaustive enumeration
- Many are instrument dependent

## Database Methods

- **False predictions** if missing protein in database
- Difficult to identify post-translational **modifications / mutations**
- Often exhibit **dependencies** on training data sets and databases

# Our Approach to Address the Peptide Identification Problem

**Novel Technique**: Using **Mixed-Integer Linear Optimization** (MILP) to formulate the peptide sequencing problem

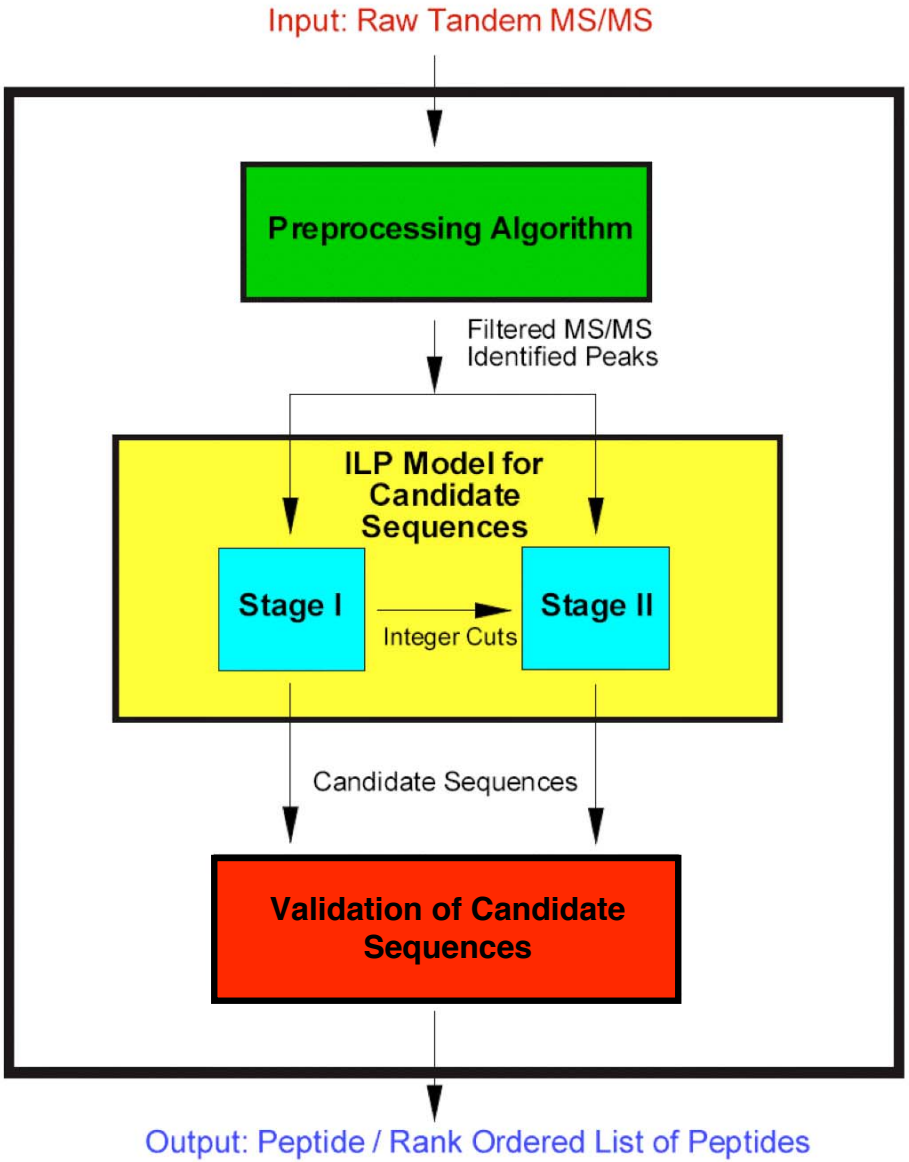



Binary variables {0-1 variables} define whether or not **peaks** ($p_i$) and **paths** between peaks ($w_{ij}$) are used in the construction of the candidate sequence, where **1** indicates **yes** and **0** indicates **no**



P.A. DiMaggio and C.A. Floudas, AIChE Journal, 53(1), 160-173 (2007).

# Algorithmic Framework




## Components of Framework:

**I. Preprocessing** of
Tandem MS Data

**II. Mathematical Model** for
Peptide Identification

**III. Postprocessing** of
Candidate Sequences

**Cross-Correlation (de novo)** **Database Alignment (hybrid)**

**De Novo: PILOT** **Hybrid: PILOT_SEQUEL**

# I. Preprocessing Algorithm

- Determine **boundary condition** ($BC^{tail}$) for the **N-terminus** of the y-ion series

- For **tryptic** peptides,

  **C-terminus** amino acid is

  **K** → 147 Da

  **R** → 175 Da

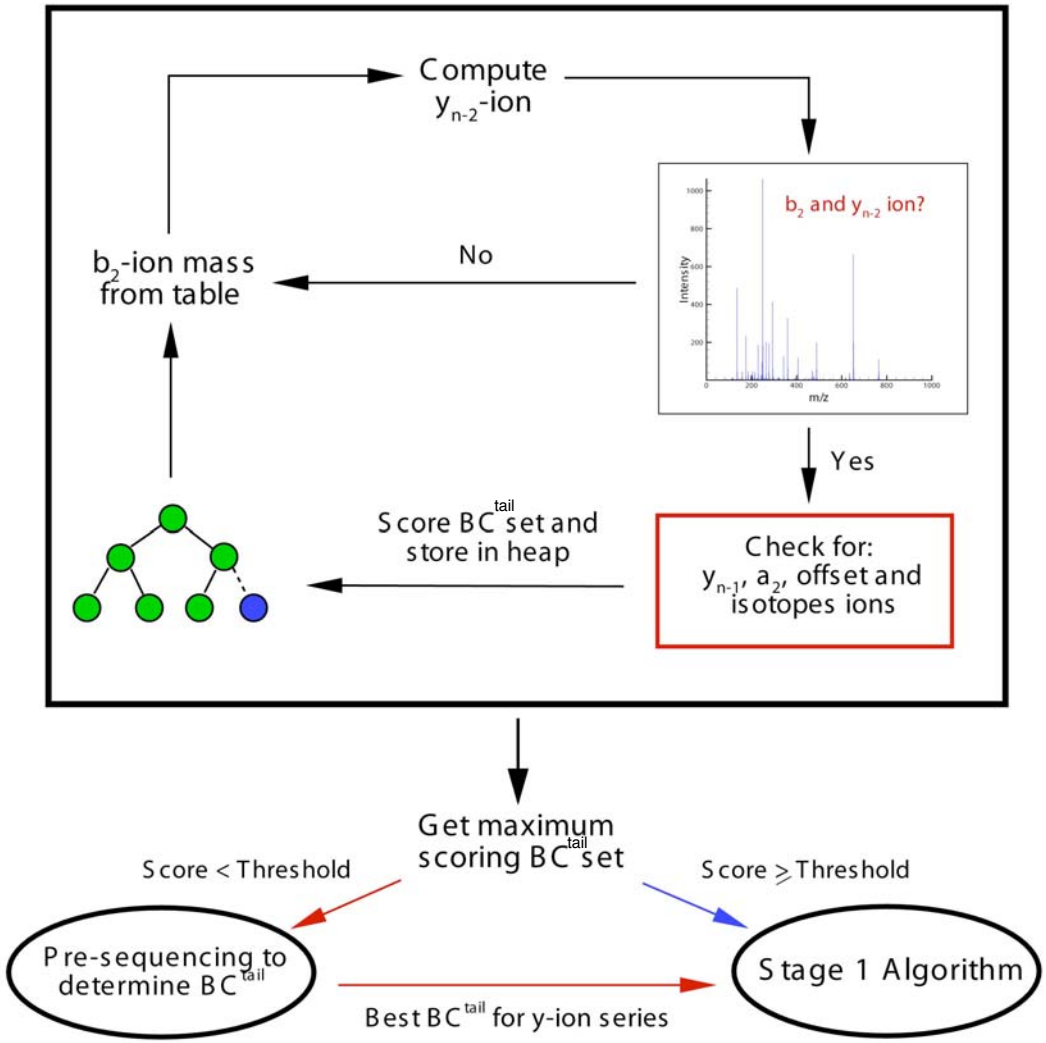
- Identify **multiply-charged** ions

  - For **high-resolution** instruments only
  - Measure distance between **isotopes**

- Identify **neutral losses** of small molecules

  i.e., -H2O, -NH3, etc.

**Algorithm for N-terminus Boundary Ions**



# II. Mathematical Model: Objective Function

$$\underset{p_k,\, w_{i,j}}{MAX} \sum_{(i,j)\in S_{i,j}} \lambda_j \cdot w_{i,j}$$

$\lambda_j$ = intensity of ion peak j

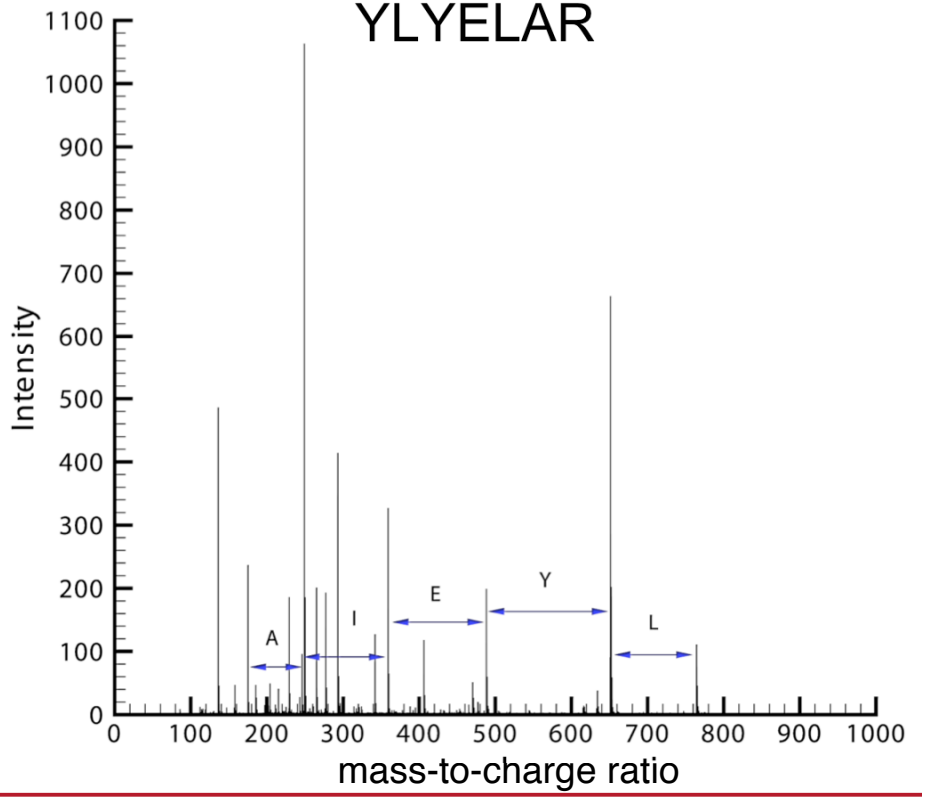
$S_{i,j} = (i,j) : m/z(\text{ion peak j}) - m/z(\text{ion peak i}) = mass(\text{amino acid})$

$$p_i = \begin{cases} 1, & \text{if peak (i) is selected} \\ 0, & \text{otherwise} \end{cases}$$

$$w_{i,j} = \begin{cases} 1, & \text{if peaks (i) and (j) are connected} \\ & \text{by a path} \quad (\text{i.e., } p_i = p_j = 1) \\ 0, & \text{otherwise} \end{cases}$$

- Maximize the use of *high intensity peaks* in constructing the candidate sequence
- Based on the observation that y- and b-ions are consistently the most abundant peaks in intensity in MS/MS

Illustration using the y-ion series for YLYELAR



# II. Mathematical Model: Constraints

## Conservation of Mass

$$\sum_{(i,j)\in S_{i,j}} M_{i,j} \cdot w_{i,j} \leq m_P + tolerance$$

$$\sum_{(i,j)\in S_{i,j}} M_{i,j} \cdot w_{i,j} \geq m_P - tolerance$$

*tolerance* "relaxes" equality

## Boundary Conditions (BC)

$$\sum_{i\in BC_i^{head}} \sum_{j\in S_{i,j}} w_{i,j} = 1$$

$$\sum_{j\in BC_j^{tail}} \sum_{i\in S_{i,j}} w_{i,j} = 1$$

- *BC* elements are dependent on ion type
- *BC* elements are checked in a preprocessing algorithm
- If elements missing then BC set is adjusted

## Complementary Ions

$$p_i + p_j \leq 1 \qquad \forall\,(i,j) \in C_{i,j}$$

b ⟷ y
a ⟷ x
c ⟷ z

Eliminates different ions of different type

# II. Mathematical Model: MILP

$$\underset{p_k,\, w_{i,j}}{MAX} \sum_{(i,j)\in S_{i,j}} \lambda_j \cdot w_{i,j}$$

$$s.t. \quad \sum_{(i,j)\in S_{i,j}} M_{i,j} \cdot w_{i,j} \leq m_P + tolerance$$

$$\sum_{(i,j)\in S_{i,j}} M_{i,j} \cdot w_{i,j} \geq m_P - tolerance$$

$$p_i + p_j \leq 1 \qquad \forall\,(i,j) \in C_{i,j}$$

Relationship between $p_i$ & $w_{i,j}$

$$\sum_{j\in S_{i,j}} w_{i,j} = p_i \qquad \forall i \in BC_i^{head}$$

$$\sum_{j\in S_{i,j}} w_{j,i} = p_i \qquad \forall i \notin BC_i^{head}$$

$$\sum_{i\in BC_i^{head}} \sum_{j\in S_{i,j}} w_{i,j} = 1$$

$$\sum_{j\in BC_j^{tail}} \sum_{i\in S_{i,j}} w_{i,j} = 1$$

*Flow conservation law*

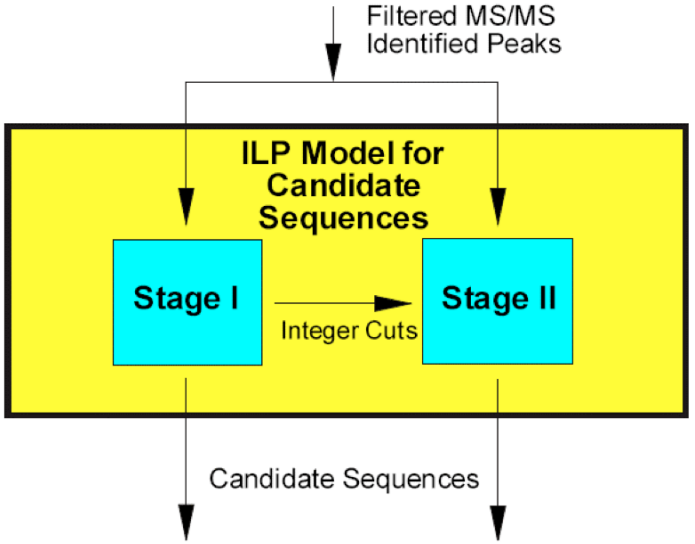
$$\sum_{j\in S_{j,i}} w_{j,i} - \sum_{k\in S_{i,k}} w_{i,k} = 0 \qquad \forall i,\ i \notin BC_i^{head},\ i \notin BC_i^{tail}$$

$$w_{i,j},\ p_k = 0 - 1 \qquad \forall (i,j),(k)$$



P.A. DiMaggio and C.A. Floudas, AIChE Journal, 53(1), 160-173 (2007).

# II. Two-Stage Framework




- During the **Stage I** calculations, the candidate sequence is constructed using only **single amino acid weights**

- Most tandem MS are **missing ion peaks** due to incomplete fragmentation and/or instruments with low m/z cutoff (i.e., ion trap mass analyzers)

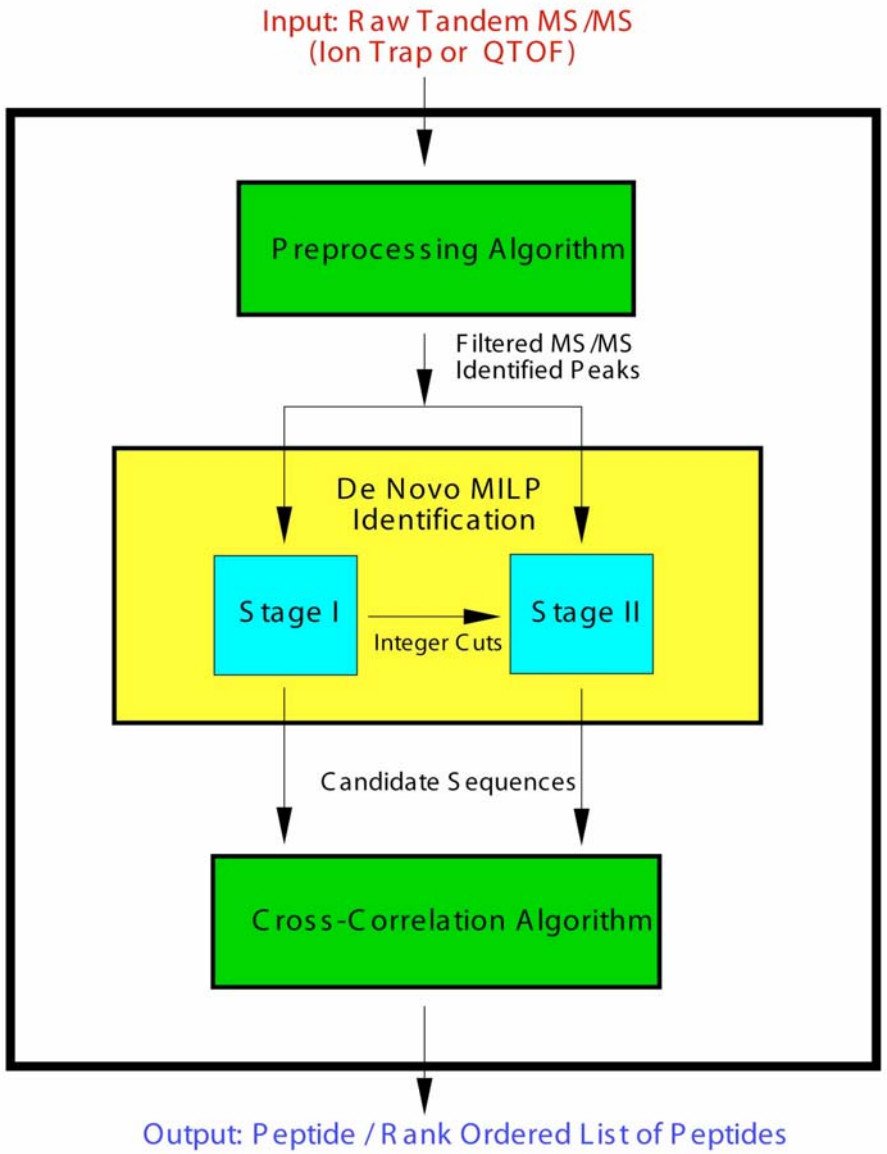
- **Stage II** calculations allow for **combinations of amino acids** to bridge the gap between missing ion peaks

- Combinations of amino acids are **penalized in objective function** to favor use of single amino acid weights in derivation of candidate sequences

# III. De Novo Post-Processing Algorithm

- Amino acid permutations substituted for **weights** in candidate sequences from Stage II calculations
- No current models exist for accurate prediction of ion **intensity trends** as a function of peptide composition for generalized mass analyzers
- Assume normalized intensity distribution + **reward** / **penalty** based on observation/absence of supporting ions
- Cross-correlate of all **theoretical** mass spectra of candidate peptide sequences with **experimental** tandem mass spectrum

# De Novo Algorithm




## Components of Framework:

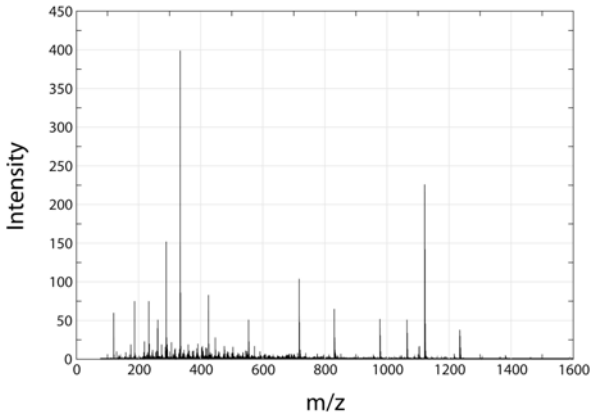
**I. Preprocessing** of Tandem MS Data

**II. Mathematical Model** for Peptide Identification

**III. Postprocessing** of Candidate Sequences

**PILOT: P**eptide identification via Mixed-**I**nteger **L**inear **O**ptimization and **T**andem mass spectrometry

# Illustrative Example for PILOT: DAFLGSFLYEYSR




Raw MS/MS spectrum

## Preprocessing Algorithm

| | |
|---|---|
| C-terminal amino acid | R → peak at 175 Da |
| N-terminus boundary conditions, $BC^{tail}$ | DA, AD, SV, VS, EG, or GE<br>no supporting $y_{n-1}$ ion |

Adjust $BC^{tail}$ → m/z($y_{n-2}$ ion) = 1381.69 Da

Filtered spectrum

Identified peaks

## Stage I Sequences

| Candidate Sequence | Objf |
|---|---|
| F(L/I)GSF(L/I)YHGANR | 2.9850 |
| F(L/I)GSF(L/I)YHAGNR | 2.9674 |
| F(L/I)GSF(L/I)YQHNR | 2.9499 |
| F(L/I)GSF(L/I)YHQNR | 2.9374 |
| F(L/I)GSF(L/I)YEYSR | 2.8547 |
| F(L/I)GSF(L/I)YEH(L/I)R | 2.7544 |
| F(L/I)GSF(L/I)YE(L/I)HR | 2.7444 |
| F(L/I)GSF(L/I)YYESR | 2.6968 |
| F(L/I)GSF(L/I)YYTDR | 2.6391 |

Integer cuts

Stage I sequences

## Stage II Sequences

| Candidate Sequence | Objf |
|---|---|
| F(L/I)GSF(L/I)Y[194.06]ANR | 2.8323 |
| F(L/I)GSF(L/I)Y[208.10]GNR | 2.8148 |
| F(L/I)GSF(L/I)Y[265.12]NR | 2.7847 |
| F(L/I)GSF(L/I)[172.04]QGANR | 2.6501 |
| F(L/I)GSF(L/I)[171.04]EGANR | 2.6501 |
| F(L/I)GSF(L/I)[171.04]SVANR | 2.6351 |
| F(L/I)GSF(L/I)[171.04]GEANR | 2.6351 |
| F(L/I)GSF(L/I)[171.04]DAANR | 2.6351 |
| F(L/I)GSF(L/I)[172.04]NAANR | 2.6351 |

All sequences

**X** = high confidence residue

**X** = low confidence residue

## PostProcessing: *DAFLGSFLYEYSR*



P.A. DiMaggio and C.A. Floudas, AIChE Journal, 53(1), 160-173 (2007).

# De Novo Comparative Study

To benchmark the performance of **PILOT**, we tested it on several tandem mass spectra from

- Quadrupole time-of-flight spectra, **QTOF** (higher resolution)
- **Ion trap** spectra (lower resolution, low m/z cutoff)

and compared the predictions to other **state-of-the-art** *de novo* methods, namely:

- **Lutefisk, LutefiskXP** – J.A. Taylor and R.S. Johnson, Anal. Chem., 73, 2594-2604 (2001).
- **PEAKS** – B. Ma et al., Rapid Commun. Mass Spec., 17, 2337-2342 (2003).
- **NovoHMM** – B. Fischer et al., Anal. Chem., 77, 7265-7273 (2005).
- **PepNovo** – A. Frank and P. Pevzner, Anal. Chem., 77, 964-973 (2005).
- **EigenMS** – M. Bern and D. Goldberg, J. Comp. Biol., 13(2), 364-378 (2006).

# De Novo Comparative Study: Ion Trap MS/MS

- **Open Proteomics Database***: contains MS/MS spectra for 5 different organisms recorded with ESI-Ion Trap mass spectrometers

- Mass spectra accompanied with predictions from SEQUEST

*Which identifications are correct?*

- Assignments examined on individual basis for quality

1. Xcorr > 2.2 and ΔCn > 0.1 for +2 charge state
2. Consistent identification with Mascot
3. $\frac{\text{Number of observed b and y ions}}{\text{Number of predicted b and y ions}}$

- Organism studied: Mycobacterium smegmatis

Xcorr = cross correlation score computed by SEQUEST

ΔCn = normalized difference in cross-correlation value between #1 and #2 hit in the search



*http://bioinformatics.icmb.utexas.edu/OPD/

# De Novo Comparative Study: Ion Trap MS/MS

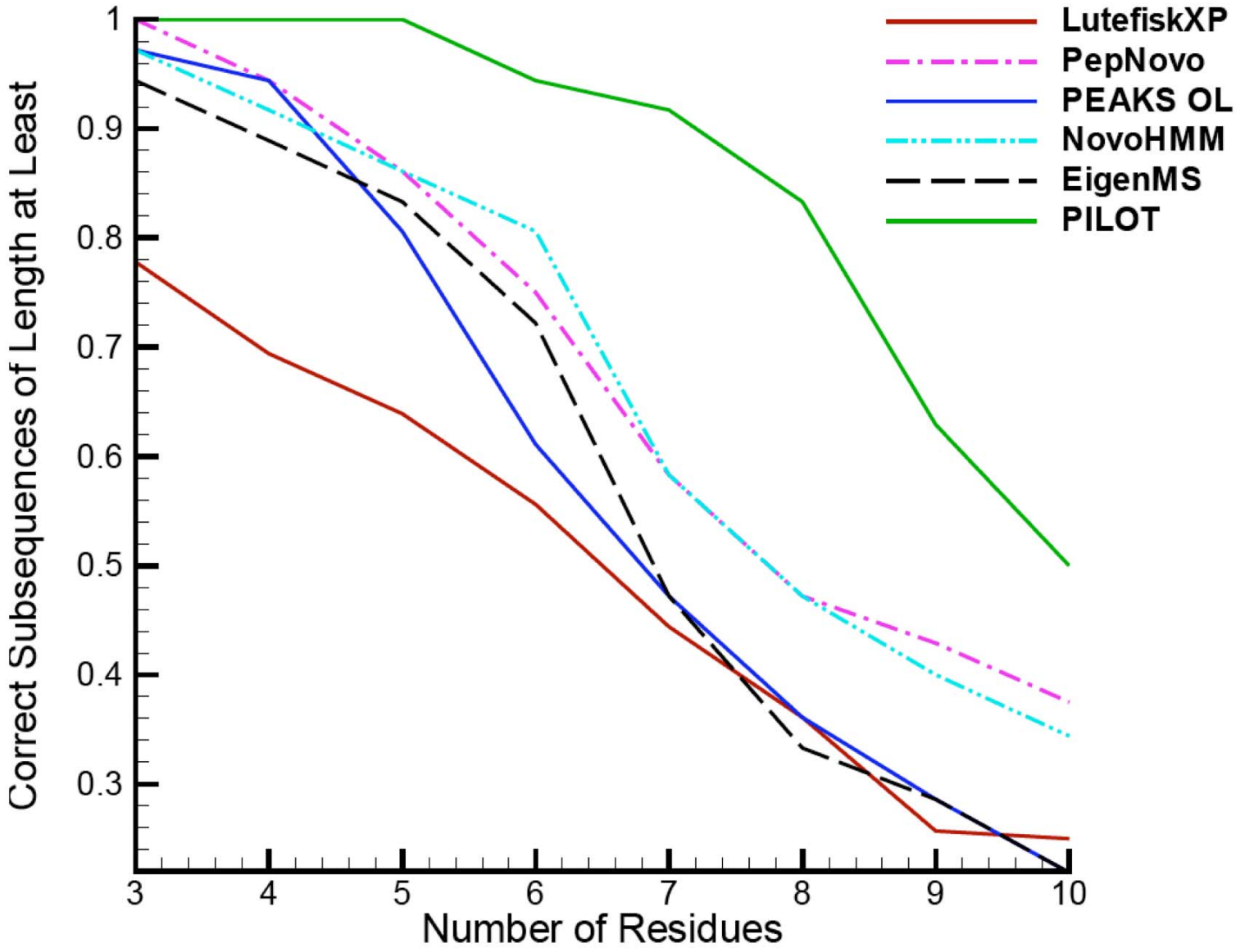
| | LutefiskXP | PepNovo | PEAKS Online | NovoHMM | EigenMS | PILOT |
|---|---|---|---|---|---|---|
| Correct Identifications | 2 (0.056) | 8 (0.222) | 6 (0.167) | 9 (0.250) | 6 (0.167) | 17 (0.472) |
| with in 1 Residue | 3 (0.083) | 9 (0.250) | 7 (0.194) | 10 (0.278) | 8 (0.222) | 17 (0.472) |
| with in 2 Residue | 11 (0.306) | 20 (0.556) | 12 (0.333) | 18 (0.500) | 18 (0.500) | 29 (0.806) |
| with in 3 Residue | 17 (0.472) | 23 (0.639) | 17 (0.472) | 25 (0.694) | 19 (0.528) | 32 (0.889) |
| Total Correct Residues | 222 (0.544) | 310 (0.760) | 281 (0.689) | 309 (0.757) | 289 (0.708) | 359 (0.880) |

| **Subsequence Length** | **x = 3** | **x = 4** | **x = 5** | **x = 6** | **x = 7** | **x = 8** | **x = 9** | **x = 10** |
|---|---|---|---|---|---|---|---|---|
| **Number of Peptides of Length ≥ x** | **36** | **36** | **36** | **36** | **36** | **36** | **35** | **32** |
| LutefiskXP | 28 (0.778) | 25 (0.694) | 23 (0.639) | 20 (0.556) | 16 (0.444) | 13 (0.361) | 9 (0.257) | 8 (0.250) |
| PepNovo | 36 (1.000) | 34 (0.944) | 31 (0.861) | 27 (0.750) | 21 (0.583) | 17 (0.472) | 15 (0.429) | 12 (0.375) |
| PEAKS Online | 35 (0.972) | 34 (0.944) | 29 (0.806) | 22 (0.611) | 17 (0.472) | 13 (0.361) | 10 (0.286) | 7 (0.219) |
| NovoHMM | 35 (0.972) | 33 (0.917) | 31 (0.861) | 29 (0.806) | 21 (0.583) | 17 (0.472) | 14 (0.400) | 11 (0.344) |
| EigenMS | 34 (0.944) | 32 (0.889) | 30 (0.833) | 26 (0.722) | 17 (0.472) | 12 (0.333) | 10 (0.286) | 7 (0.219) |
| PILOT | 36 (1.000) | 36 (1.000) | 36 (1.000) | 34 (0.944) | 33 (0.917) | 30 (0.833) | 22 (0.629) | 16 (0.500) |



P.A. DiMaggio and C.A. Floudas, Anal. Chem., 79, 1433-1446 (2007).

# De Novo Comparative Study: Ion Trap MS/MS






P.A. DiMaggio and C.A. Floudas, Anal. Chem., 79, 1433-1446 (2007).

# De Novo Comparative Study: QTOF MS/MS

- *Quadrupole time-of-flight* (QTOF) spectra have better resolution that ion trap spectra
- Examined QTOF data for a mixture of 4 known proteins*:
  - ✓ Alcohol dehydrogenase (yeast)
  - ✓ Myoglobin (horse)
  - ✓ Albumin (horse, BSA)
  - ✓ Cytochrome C (horse)
- Spectra were assessed for quality based on the metric:

$$\frac{s}{m} = \frac{\sum_{\{i\,:\lambda_i > 2\}} \lambda_i}{\text{Peptide Mass}}$$

($\lambda_i$ = intensity of ion peak i)



*http://www.csd.uwo.ca/~bma/peaks/

# De Novo Comparative Study: QTOF MS/MS

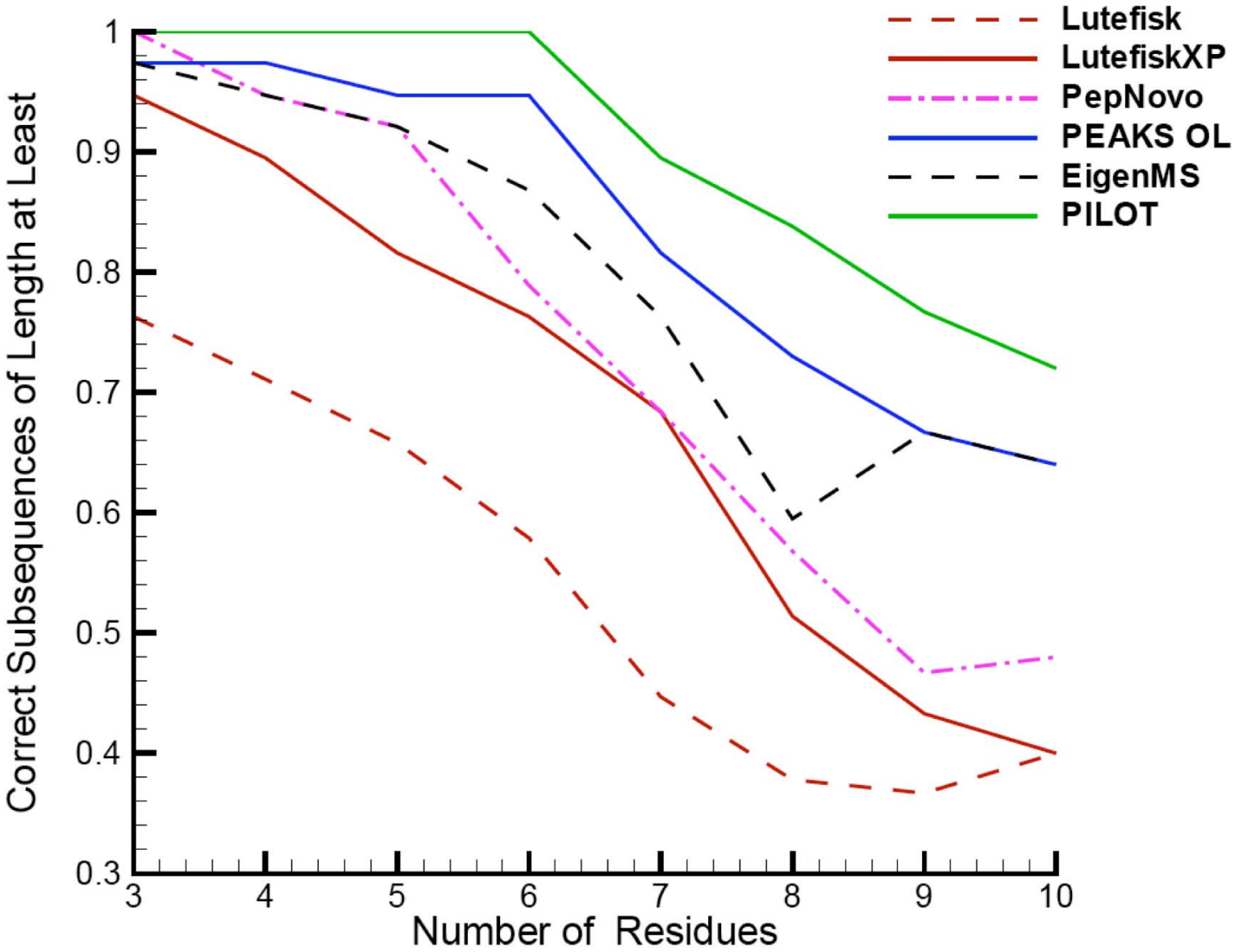
| | Lutefisk | LutefiskXP | PepNovo | PEAKS Online | EigenMS | PILOT |
|---|---|---|---|---|---|---|
| Correct Identifications | 10 (0.263) | 9 (0.237) | 16 (0.421) | 21 (0.553) | 20 (0.526) | 25 (0.658) |
| with in 1 Residue | 11 (0.290) | 10 (0.263) | 17 (0.447) | 22 (0.579) | 21 (0.553) | 25 (0.658) |
| with in 2 Residue | 23 (0.605) | 22 (0.579) | 25 (0.658) | 29 (0.763) | 29 (0.763) | 33 (0.868) |
| with in 3 Residue | 23 (0.605) | 25 (0.658) | 27 (0.711) | 32 (0.842) | 30 (0.790) | 35 (0.921) |
| Total Correct Residues | 245 (0.586) | 294 (0.703) | 337 (0.806) | 366 (0.876) | 353 (0.845) | 381 (0.912) |

| **Subsequence Length** | **x = 3** | **x = 4** | **x = 5** | **x = 6** | **x = 7** | **x = 8** | **x = 9** | **x = 10** |
|---|---|---|---|---|---|---|---|---|
| **Number of Peptides of Length ≥ x** | **38** | **38** | **38** | **38** | **38** | **37** | **30** | **25** |
| Lutefisk | 29 (0.763) | 27 (0.711) | 25 (0.658) | 22 (0.579) | 17 (0.447) | 14 (0.378) | 11 (0.367) | 10 (0.400) |
| LutefiskXP | 36 (0.947) | 34 (0.895) | 31 (0.816) | 29 (0.763) | 26 (0.684) | 19 (0.514) | 13 (0.433) | 10 (0.400) |
| PepNovo | 38 (1.000) | 36 (0.947) | 35 (0.921) | 30 (0.789) | 26 (0.684) | 21 (0.568) | 14 (0.467) | 12 (0.480) |
| PEAKS Online | 37 (0.974) | 37 (0.974) | 36 (0.947) | 36 (0.947) | 31 (0.816) | 27 (0.730) | 20 (0.667) | 16 (0.640) |
| EigenMS | 37 (0.974) | 36 (0.947) | 35 (0.921) | 33 (0.868) | 29 (0.763) | 22 (0.595) | 20 (0.667) | 16 (0.640) |
| PILOT | 38 (1.000) | 38 (1.000) | 38 (1.000) | 38 (1.000) | 34 (0.895) | 31 (0.838) | 23 (0.767) | 18 (0.720) |



P.A. DiMaggio and C.A. Floudas, Anal. Chem., 79, 1433-1446 (2007).

# De Novo Comparative Study: QTOF MS/MS






P.A. DiMaggio and C.A. Floudas, Anal. Chem., 79, 1433-1446 (2007).

# De Novo Method Summary

- Developed accurate **de novo** framework, **PILOT**, for the **identification of peptides** via tandem mass spectrometry (MS/MS)
- **PILOT** outperformed several state-of-the-art de novo methods in a **comparative study** for ion trap and QTOF tandem mass spectra
- **Key elements** of de novo framework:
  - Novel mixed-integer linear optimization (MILP) formulation for peptide identification
  - Preprocessing algorithm for filtering spectra and identifying important ion peaks
  - Post-processing algorithm for cross-correlating theoretical tandem mass spectra with experimental tandem mass spectrum

# Hybrid Method for Peptide Identification

- Main Idea: Can use **protein databases** to resolve **ambiguous residue assignments** from *de novo* sequence predictions
- Combine strengths of **de novo** and **database** methods

```
361 HPEYAVSVLL RLAKEYEATL EDCCAKEDPH ACYATVFDKL
    HPEYAVEGLL R
```

- Local database search tools, such as **FASTA***, can be utilized to align *de novo* sequences in a protein database

... but several modifications are necessary



*W. Pearson and D. Lipman. *PNAS*, 85:2444-2448, 1988.

# II. Modified ILP Model for Hybrid Method

$$\underset{p_k,\, w_{i,j}}{MAX} \left( \frac{1}{\omega} \sum_{i \in S_{i,j}} \lambda_i \cdot w_{i,j} + \sum_{i \in C_{i,j'}, S_{i,j}} (\lambda_i + \lambda_{j'}) \cdot p_i \right)$$

$$s.t. \quad \sum_{(i,j) \in S_{i,j}} M_{i,j} \cdot w_{i,j} \leq \sum_{i \in BC_i^{tail}} m/z_i \cdot p_i + tolerance$$

$$\sum_{(i,j) \in S_{i,j}} M_{i,j} \cdot w_{i,j} \geq \sum_{i \in BC_i^{tail}} m/z_i \cdot p_i - tolerance$$

*Conservation of Mass relaxed by tolerance*

*Complementary Ions* →

$$p_i + p_j \leq 1 \qquad \forall\, (i,j) \in C_{i,j}$$

*Relationship between $p_i$ & $w_{i,j}$*

$$\sum_{j \in S_{i,j}} w_{i,j} = p_i \qquad \forall i \in BC_i^{head}$$

$$\sum_{j \in S_{j,i}} w_{j,i} = p_i \qquad \forall i \notin BC_i^{head}$$

$$\sum_{i \in BC_i^{head}} \sum_{j \in S_{i,j}} w_{i,j} = 1$$

$$\sum_{j \in BC_j^{tail}} \sum_{i \in S_{i,j}} w_{i,j} = 1$$

*Boundary Conditions for C-terminus and N-terminus*

*Tryptic Peptide* →

$$\sum_{i \in TP_i} p_i = 1$$

*Flow conservation law*

$$\sum_{j \in S_{j,i}} w_{j,i} - \sum_{k \in S_{i,k}} w_{i,k} = 0 \qquad \forall i,\ i \notin BC_i^{head},\ i \notin BC_i^{tail}$$

$$w_{i,j},\ p_k = \{0,1\} \qquad \forall (i,j),(k)$$

# III. PostProcessing using Modified FASTA Algorithm

## ❑ Scoring Matrices

Modify **BLOSUM** matrix to conserve mass between query and template sequences

| | A | R | N | D | C | Q | E | G | H | I | L | K | M | F | P | S | T | W | Y | V | B | Z | X |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | 1 |
| R | -5 | 5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | 1 |
| N | -5 | -5 | 5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | 1 |
| D | -5 | -5 | -5 | 5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | 1 |
| C | -5 | -5 | -5 | -5 | 5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | 1 |
| Q | -5 | -5 | -5 | -5 | -5 | 5 | -5 | -5 | -5 | -5 | -5 | 5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | 1 |
| E | -5 | -5 | -5 | -5 | -5 | -5 | 5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | 1 |
| G | -5 | -5 | -5 | -5 | -5 | -5 | -5 | 15 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | 1 |
| H | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | 5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | 1 |
| I | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | 5 | 5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | 1 |
| L | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | 5 | 5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | 1 |
| K | -5 | -5 | -5 | -5 | -5 | 5 | -5 | -5 | -5 | -5 | -5 | 5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | 1 |
| M | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | 5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | 1 |
| F | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | 5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | 1 |
| P | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | 15 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | 1 |
| S | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | 5 | -5 | -5 | -5 | -5 | -5 | -5 | 1 |
| T | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | 5 | -5 | -5 | -5 | -5 | -5 | 1 |
| W | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | 5 | -5 | -5 | -5 | -5 | 1 |
| Y | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | 5 | -5 | -5 | -5 | 1 |
| V | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | 5 | -5 | -5 | 1 |
| B | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | 5 | -5 | 1 |
| Z | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | -5 | 5 | 1 |
| X | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |

## ❑ Hashing

***ktup = 4*** to optimize only high quality sequence matches

## ❑ Smith and Waterman Optimization

- **Isobaric** residues

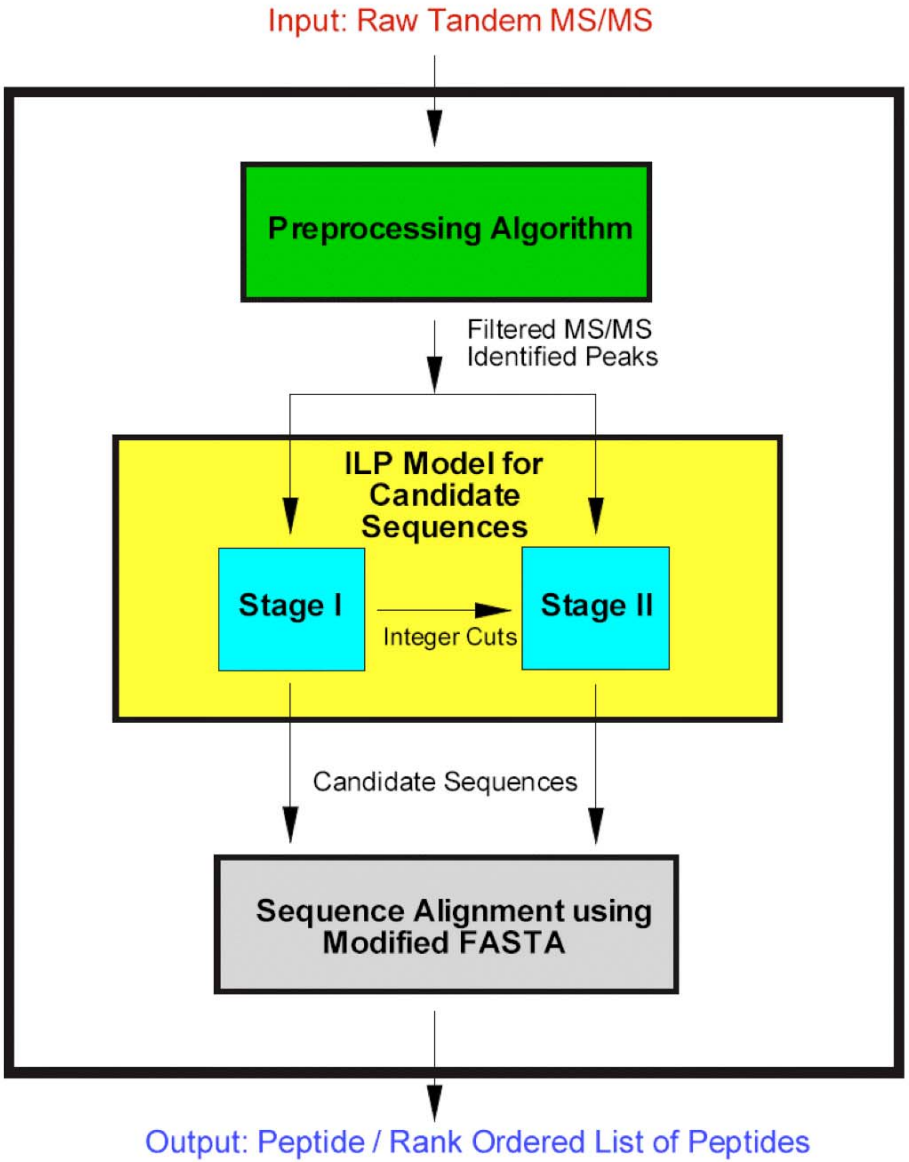

- Explicit **Conservation of Mass** between template & query

## ❑ Tryptic Peptide Databases



P.A. DiMaggio, C.A. Floudas, B. Lu, and J.R Yates, *J. Proteome Res.*, 7, 1584-1593 (2008).

# Hybrid Algorithm




## Components of Framework:

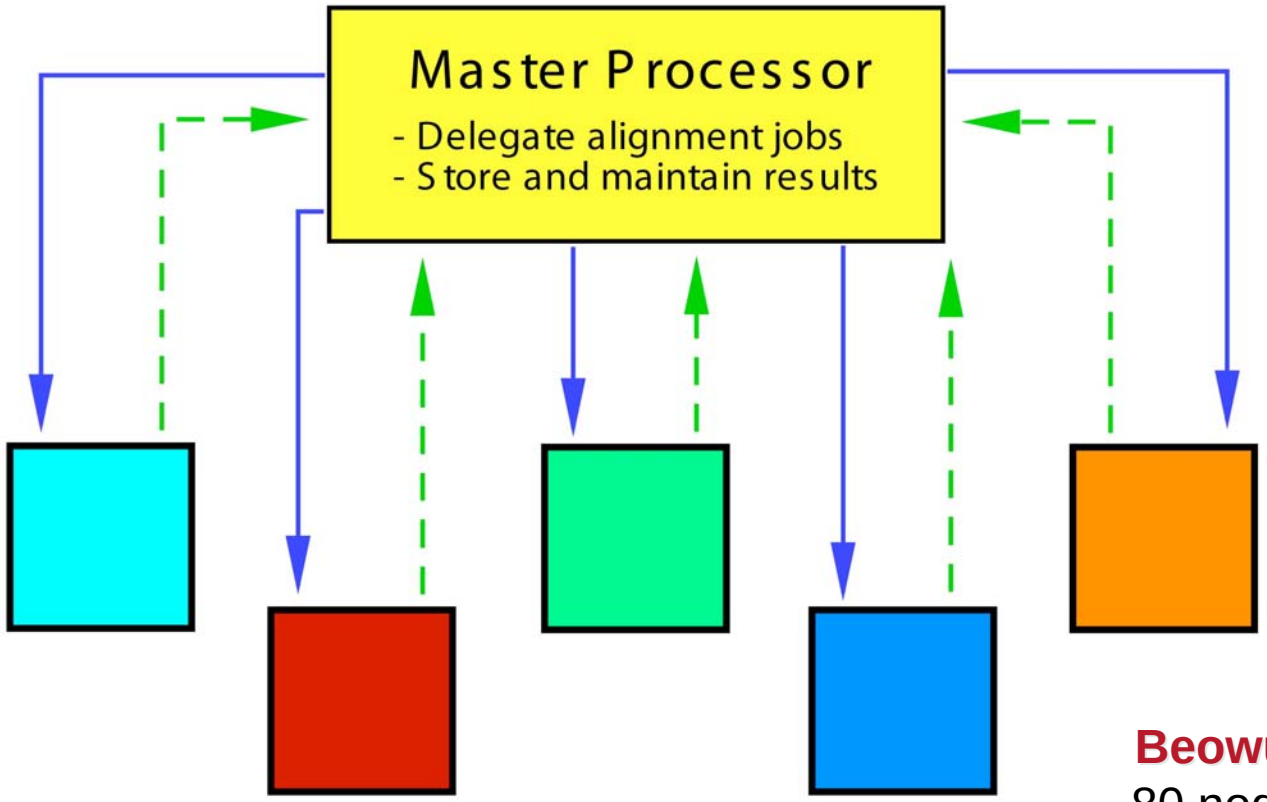
**I. Preprocessing** of Tandem MS Data

**II. Mathematical Model** for Peptide Identification

**III. Postprocessing** of Candidate Sequences

**PILOT_SEQUEL: P**eptide identification via Mixed-**I**nteger **L**inear **O**ptimization, and **T**andem mass spectrometry, and local **SEQUE**nce a**L**ignment

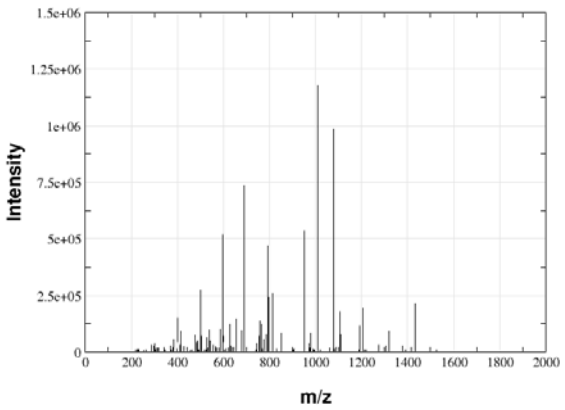
# Hybrid Approach for Peptide Identification
## Distributive Computing Framework




**Beowulf Cluster** :
80 nodes with dual Intel Xeon 3.0 GHz processors

**PILOT_SEQUEL: P**eptide identification via Mixed-**I**nteger **L**inear **O**ptimization, and **T**andem mass spectrometry, and local **SEQUE**nce a**L**ignment

# Example for PILOT_SEQUEL: VEADIAGHGQEVLIR




Raw MS/MS spectrum

## Preprocessing Algorithm

| | |
|---|---|
| C-terminal amino acid | Peaks observed at 147 and 175 Da → allowed to be both K and R |
| N-terminus boundary conditions, $BC^{tail}$ | No N-terminal pair with significant score: Upper bounding calculations select 1192 Da and 1378 Da as N-terminal boundary conditions |

Filtered spectrum

N- and C-terminal boundary conditions

## Stage I Sequences

| Candidate Sequence | Objf |
|---|---|
| **IAG**HG**QEVI**IR | 3.43 |
| ANN**AG**HG**QEVI**IR | 3.38 |
| **IAG**HGWA**VI**IR | 3.32 |
| ANN**AG**HGWA**VI**IR | 3.26 |
| **IAG**HG**Q**QVNIR | 3.04 |
| ANN**AG**HG**Q**QVNIR | 2.98 |
| **IA**CFE**EVI**IR | 2.78 |
| ANN**A**CFE**EVI**IR | 2.73 |
| AT**VVG**HG**Q**QVNIR | 2.41 |

**bold font** = high confidence residues

Integer cuts

Stage I sequences

## Stage II Sequences

| Candidate Sequence | Objf |
|---|---|
| **IAG[194.05]QEVI**IR | 3.41 |

Query all sequences in nr protein database

De novo peptide: **VEW IAGHGQEVLIR**

Database protein: LKTEAEMK**VEADLAGHGQEVLIR**

**Selected Peptide: *VEADIAGHGQEVLIR***

# Hybrid Comparative Study: OrbiTrap MS/MS

- **OrbiTrap** instruments have 2-3 times the resolution of conventional mass spectrometers.
- Examined 380 OrbiTrap tandem MS for a control mixture of 16 **known proteins** from:
  - ✓ *bovine, bovine serum, horse, chicken, rabbit, ecoli*
- **SEQUEST** was used to search a **target protein database** comprised of 5009 proteins (appended with a database containing the *reversed* sequences of these proteins).
- The validity of the spectra/peptide matches were assessed using **DTASelect***.



*D.L. Tabb et al. *J. Proteome Res.*, 1:21-26, 2002.

# Hybrid Comparative Study: OrbiTrap MS/MS

| | Mascot | CIDentify (PepNovo) | PepNovo* | InsPecT, InsPecT L=6 | CIDentify (PILOT) | PILOT_SEQUEL |
|---|---|---|---|---|---|---|
| Correct Identifications | 286 (0.753) | 287 (0.755) | 292 (0.768) | 280 (0.737), 264 (0.695) | 298 (0.784) | 352 (0.926) |
| with in 1 Residue | 287 (0.755) | 287 (0.755) | 292 (0.768) | 294 (0.774), 267 (0.703) | 299 (0.787) | 352 (0.926) |
| with in 2 Residue | 289 (0.760) | 291 (0.766) | 296 (0.780) | 351 (0.924), 322 (0.847) | 313 (0.824) | 356 (0.936) |
| with in 3 Residue | 289 (0.760) | 291 (0.766) | 298 (0.784) | 352 (0.926), 323 (0.850) | 318 (0.837) | 357 (0.939) |
| Total Correct Residues | 3638 (0.834) | 3544/4364 (0.812) | 3564/4364 (0.820) | 4045 (0.927), 3806 (0.872) | 3841 (0.880) | 4159 (0.953) |

*De novo algorithm trained on OrbiTrap tandem MS[+]. A sequence tag algorithm was used to perform database search in place of the direct lookup method based on hashing.

**Residues predicted** for de novo sequences:

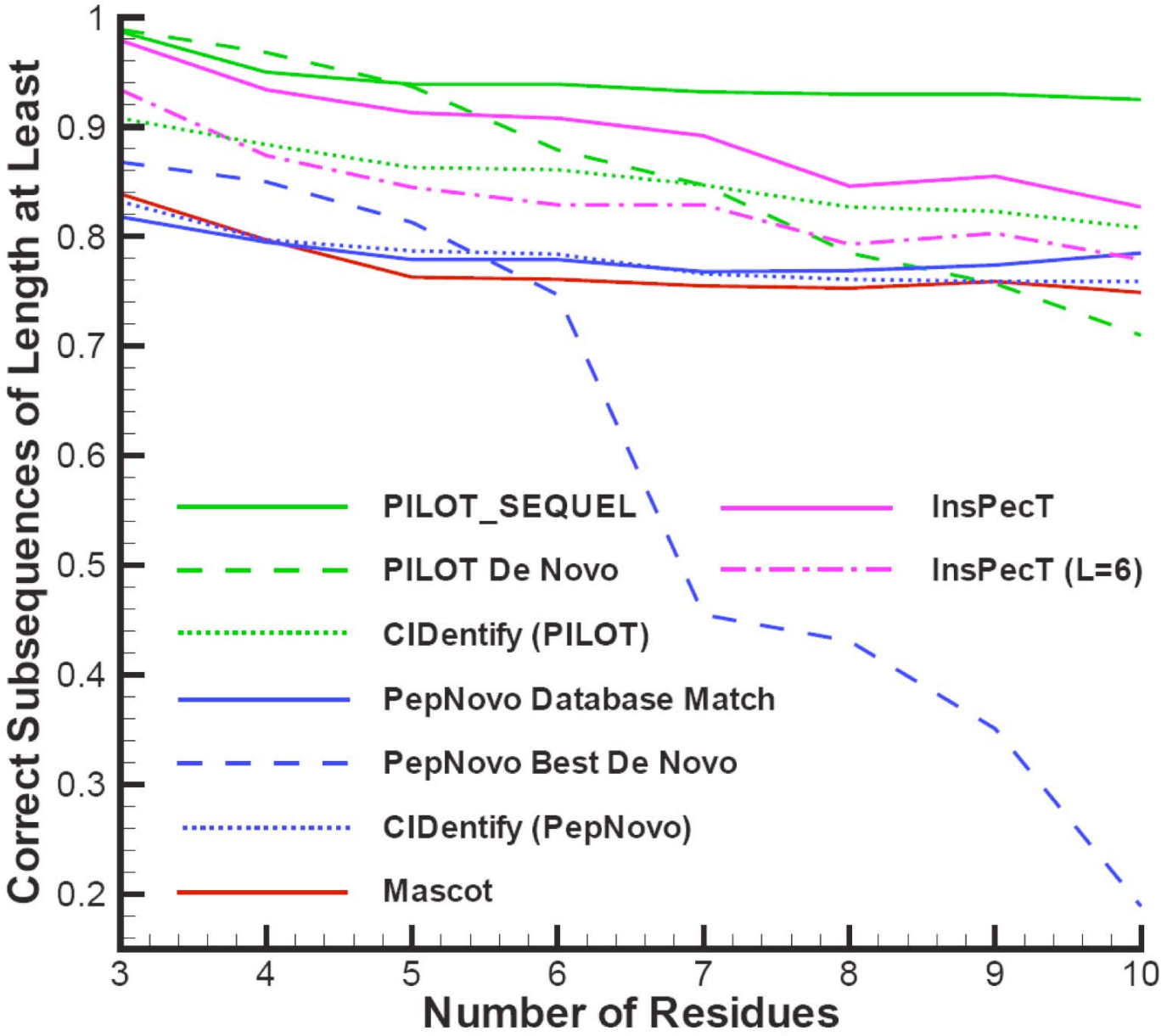
PILOT 4249/4364 (0.974)

PepNovo 2958/4364 (0.678)



[+]A. Frank et al. *J. Proteome Res.*, 6:114-123, 2007.

# Hybrid Comparative Study: OrbiTrap MS/MS






P.A. DiMaggio, C.A. Floudas, B. Lu, and J.R Yates, *J. Proteome Res.*, 7, 1584-1593 (2008).

# Hybrid Method Summary

- Developed accurate **hybrid** framework, for the **identification of peptides** via tandem mass spectrometry (MS/MS) which combines strengths of **de novo and database** techniques

- **PILOT_SEQUEL** outperformed several state-of-the-art algorithms for **hybrid** and **database** peptide identification in a **comparative study** using OrbiTrap tandem mass spectra.

- Major components of hybrid method:
  - Modified integer linear optimization (ILP) formulation for peptide identification
  - Enhanced implementation of FASTA
  - Distributed computing framework for performing several sequence alignment calculations