PhD Thesis Proposal

Conditional Graphical Models for Protein Structure Prediction

Yan Liu
Language Technologies Institute
Carnegie Mellon University

Thesis Committee

Jaime Carbonell (Chair)
John Lafferty
Eric P. Xing
Vanathi Gopalakrishnan (Univ. of Pittsburgh)
Biology on One Slide

Predict protein structures from sequences
Protein Structure Hierarchy

<table>
<thead>
<tr>
<th>Primary Structure</th>
<th>Secondary Structures</th>
<th>Tertiary Structures</th>
<th>Quaternary Structures</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image" alt="Primary Structure Diagram" /></td>
<td><img src="image" alt="Secondary Structures Diagram" /></td>
<td><img src="image" alt="Tertiary Structures Diagram" /></td>
<td><img src="image" alt="Quaternary Structures Diagram" /></td>
</tr>
</tbody>
</table>

Our focus: predicting the topology of the structures from sequences

APAFSVSPASGACGPECA
Previous work

• General approaches
  ➢ Sequence similarity searches, e.g. PSI-BLAST [Altschul et al, 1997]
  ➢ Profile HMM, e.g. HMMER [Durbin et al, 1998] and SAM [Karplus et al, 1998]
  ➢ Homology modeling or threading, e.g. Threader [Jones, 1998]
  ➢ Window-based methods, e.g. PSI_pred [Jones, 2001]

Fail to handle
- Structural similarity without clear sequence similarity
- Long-range interactions

• Methods of careful design for specific structures
  ➢ Example: αα- and ββ- hairpins, β-turn and β-helix [Efimov, 1991; Wilmot and Thornton, 1990; Bradley at al, 2001]

Hard to generalize
Missing pieces

• No principled probabilistic models to formulate the structured properties of proteins

  Graphical models

• Informative features without clear mapping to structures
  ➢ Polar, hydrophobic, aromatic and etc..

  Discriminative models
Thesis Statement

Conditional graphical models are effective for protein structure prediction
Conditional Graphical Models (CGM)

• Discriminative model defined over undirected graphs
  ➢ Conditional random fields [Lafferty et al, 2001]

• Protein structural graph
  ➢ Nodes: structural building blocks or observed sequences
  ➢ Edges: interactions between blocks (local or long-range)

• Structural prediction is reduced to segmentation and labeling problem
Conditional Graphical Models (CGM)

- Define segment semantics as structural building blocks
  - \( W = (M, \{W_i\}) \)
    - \( M \): # of segments (predetermined or to be inferred)
    - \( W_i \): a set of labels determining segment \( i \)

- The conditional probability of a possible segmentation \( W \) given the observation \( x \) is defined as
  \[
P(W \mid x) = \frac{1}{Z} \prod_{c \in C_G} \exp\left( \sum_{k=1}^{K} \lambda_k f_k(x, w_c) \right)
  \]
Advantages of Conditional Graphical Models

- Flexible feature definition and kernelization
- Different loss functions and regularizers
- Convex functions for globally optimal solutions
- Structured information, especially the long-range interactions
Training and Testing

• Training phase: learn the model parameters
  
  ➢ Minimizing regularized loss function
  
  \[
  L_\lambda = \sum_{c \in C, k = 1}^{K} \lambda_k f_k(x, W_c) - \log Z + \Omega(\|\lambda\|)
  \]

  ➢ Seek the direction whose empirical values agree with the expectation

  \[
  \frac{\partial L_\lambda}{\partial \lambda_k} = \sum_{c \in C, \lambda} (f_k(x, W_c) - E_{p(W|x)}[f_k(x, W_c)]) + \nabla \Omega(\|\lambda\|) = 0
  \]

  ➢ Iterative searching algorithms have to be applied

• Testing phase: search the segmentation that maximizes \(P(w \mid x)\)
## Thesis Work

Conditional graphical models for protein structure prediction

<table>
<thead>
<tr>
<th>Hierarchy</th>
<th>Secondary Structures</th>
<th>Tertiary Structures</th>
<th>Quaternary Structures</th>
</tr>
</thead>
<tbody>
<tr>
<td>Examples</td>
<td><img src="image1" alt="Example" /></td>
<td><img src="image2" alt="Example" /></td>
<td><img src="image3" alt="Example" /></td>
</tr>
<tr>
<td>Structural building blocks</td>
<td>Amino acids (residues)</td>
<td>Secondary structure elements</td>
<td>Secondary structure and tertiary structure elements</td>
</tr>
<tr>
<td>Specific model</td>
<td>CRFs Kernel CRFs</td>
<td>Segmentation CRFs Chain graph model</td>
<td>Factorial Segmentation CRFs</td>
</tr>
<tr>
<td>Status</td>
<td>Completed</td>
<td>Work in progress</td>
<td>Proposed work</td>
</tr>
</tbody>
</table>
Model Roadmap

Conditional random fields → Kernelized → Kernel CRFs

Segmentation CRFs → Locally normalized → Chain graph model

Factorized → Generalized conditional graphical models

Factorial segmentation CRFs
## Outline

<table>
<thead>
<tr>
<th>Hierarchy</th>
<th>Secondary Structures</th>
<th>Tertiary Structures</th>
<th>Quaternary Structures</th>
</tr>
</thead>
<tbody>
<tr>
<td>Examples</td>
<td><img src="image1" alt="Examples" /></td>
<td><img src="image2" alt="Examples" /></td>
<td><img src="image3" alt="Examples" /></td>
</tr>
<tr>
<td>Structural building blocks</td>
<td>Amino acids (residues)</td>
<td>Secondary structure elements</td>
<td>Secondary structure and tertiary structure elements</td>
</tr>
<tr>
<td>Specific model</td>
<td>CRFs</td>
<td>Segmentation CRFs</td>
<td>Factorial Segmentation CRFs</td>
</tr>
<tr>
<td></td>
<td>Kernel CRFs</td>
<td>Chain graph model</td>
<td></td>
</tr>
<tr>
<td>Status</td>
<td>Completed</td>
<td>Work in progress</td>
<td>Proposed work</td>
</tr>
</tbody>
</table>
Task Definition and Evaluation Materials

• Given a protein sequence, predict its secondary structure assignments
  ➢ Three class: helix (C), sheets (E) and coil (C)

Protein Secondary Structure Prediction
CGM on Secondary Structure Prediction


• Structural component
  ➢ Individual residues

• Segmentation definition
  ➢ $W = (n, Y)$, $n$: # of residues, $Y_i = H, E$ or $C$

• Specific models
  ➢ Conditional random fields (CRFs) [lafferty et al, 2001]
  ➢ Kernel conditional random fields (kCRFs)
    • where
      $$ f^* = \sum_{j=1}^{L} \sum_{c \in C_{G_{j}}} \sum_{y_c \in \mathcal{Y}} \lambda_{y_c} K(\bullet, (x^{(j)}, c, y_c)) $$

$$ P(y \mid x) = \frac{1}{Z} \prod_{i=1}^{N} \exp \left( \sum_{k=1}^{K} \lambda_k f_k(x, i, y_{i-1}, y_i) \right) $$

$$ P(y \mid x) = \frac{1}{Z} \prod_{c \in C_G} \exp \left( f^*(x, c, y_c) \right) $$
Lessons from Previous Work

- How to combine the predictions effectively?
- How to accurately identify beta-sheets?
- How to explore evolutionary features via kernels?
Experiment Results on Prediction Combination

- Previous work:
  - Window-based label combination [Rost and Sander, 1993]
  - Window-based score combination [Jones, 1999]
- Other graphical models for score combination
  - Maximum entropy Markov model (MEMM) [McCallum et al, 2000]
  - Higher-order MEMMs (H-MEMM), Pseudo state duration MEMMs (PSMEMM)

- Graphical models are consistently better than the window-based approaches
- CRFs perform the best among the four graphical models
Experiment Results on Beta-sheet Detection

- Offset from the correct register for the correctly predicted beta-strand pairs (CB513 dataset)

Prediction accuracy for beta-sheets

Considerable improvement in prediction performance
Protein Secondary Structure Prediction

Experiment Results on Feature Exploration

- Prediction accuracy for KCRFs with vertex cliques (V) and edge cliques (E)
  - PSI-BLAST profile with RBF kernel

![Accuracy (Q3) and State Transition Accuracy charts](chart.png)
Summary

- Conditional graphical model for protein secondary structure prediction
  - Conditional random fields (CRFs)
  - Kernel conditional random fields (KCRFs)
Outline

Conditional graphical models for protein structure prediction

<table>
<thead>
<tr>
<th>Hiarchy</th>
<th>Secondary Structures</th>
<th>Tertiary Structures</th>
<th>Quaternary Structures</th>
</tr>
</thead>
<tbody>
<tr>
<td>Examples</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Structural building blocks</td>
<td>Amino acids (residues)</td>
<td>Secondary structure elements</td>
<td>Secondary structure and tertiary structure elements</td>
</tr>
<tr>
<td>Specific model</td>
<td>CRFs</td>
<td>Segmentation CRFs</td>
<td>Factorial Segmentation CRFs</td>
</tr>
<tr>
<td>Status</td>
<td>Completed</td>
<td>Work in progress</td>
<td>Proposed work</td>
</tr>
</tbody>
</table>
Structural motif recognition

Task Definition and Evaluation Materials

• Structural motif
  ➢ Regular arrangement of secondary structural elements
  ➢ Super-secondary structure, or protein fold

..APAFSVSPASGACGPECA..

↓

Contains the structural motif?

Yes

..NNEEEECCCCCHHHCCC..
Carnegie Mellon School of Computer Science

CGM for Structural Motif Recognition

- Structural component
  - Secondary structure elements

- Protein structural graph
  - Nodes for the states of secondary structural elements of unknown lengths
  - Edges for interactions between nodes in 3-D

- Example: β-α-β motif
Carnegie Mellon School of Computer Science

CGM for Structural Motif Recognition

[Liu, Carbonell, Weigele and Gopalakrishnan, RECOMB 2005]

- Segmentation definition
  - \( Y_i \): state of segment \( i \)
  - \( S_i \): the length of segment \( i \)
  - \( M \): number of segments

- Segmentation conditional random fields (SCRFs)
  - For any graph, we have
    \[
    P(W | x) = \frac{1}{Z} \prod_{c \in C_G} \exp(\sum_{k=1}^{K} \lambda_k f_k(x, w_c))
    \]
  - For a simplified graph of chains or trees, we have
    \[
    P(W | x) = \frac{1}{Z} \exp(\sum_{i=1}^{M} \sum_{k=1}^{K} \lambda_k f_k(x, W_i, W_{\pi_i}))
    \]
Protein Folds with Structural Repeats

- Prevalent in proteins and important in functions
- Each rung consists of structural motifs and insertions
- Challenge
  - Low sequence similarity in structural motifs
  - Long-range interactions
Structural Motif Recognition

CGM for Structural Motif Recognition

[Liu, Xing and Carbonell, Submitted 2005]

• Chain graph
  ➢ A graph consisting of directed and undirected graphs
  ➢ Given a variable set V that forms multiple subgraphs $U$
  \[ P(V) = \prod_{u \in U} P(u \mid \text{parents}(u)) \]

• Segmentation definition
  ➢ Envelope: one rung with motifs and insertions
  ➢ Two layer segmentation $W = \{M, \{\Xi_i\}, T\}$
    M: # of envelopes
    $\Xi_i$: the segmentation of envelopes
    $T_i$: the state of envelope $i = \text{repeat, non-repeat}$

• Chain graph model
  \[
P(W \mid x) = P(M, \{\Xi_i\}, T \mid x) = P(M) \prod_{i=1}^{M} P(T_i \mid x, T_{i-1}, \Xi_{i-1}) P(\Xi_i \mid x, T_i, \Xi_{i-1})
  \]

Motif profile model  SCRFs
Experiments

• Right-handed $\beta$-helix fold
  - An elongated helix-like structures whose successive rungs composed of three parallel $\beta$-strands (B1, B2, B3 strands)
  - T2 turn: a unique two-residue turn
  - Low sequence identity (under 25%)

• Leucine-rich repeats (LLR)
  - Solenoid-like regular arrangement of beta-strands and an alpha-helix, connected by coils
  - Relatively high sequence identity (many Leucines)
Experiment Results on SCRF for $\beta$-helix

- Cross-family validation for classifying $\beta$-helix proteins
  - SCRFs can score all known $\beta$-helices higher than non $\beta$-helices

<table>
<thead>
<tr>
<th>SCOP family</th>
<th>PDB-id</th>
<th>Struct HMMs Rank</th>
<th>Seq HMMs Rank</th>
<th>Threader Rank</th>
<th>BetaWrap$^+$ Score Rank</th>
<th>SCRFs $\rho$-score Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>P.69 pertactin</td>
<td>1dab</td>
<td>3</td>
<td>75</td>
<td>24</td>
<td>-17.84</td>
<td>1</td>
</tr>
<tr>
<td>Chondroitinase B</td>
<td>1dbg</td>
<td>5</td>
<td>55</td>
<td>47</td>
<td>-19.55</td>
<td>1</td>
</tr>
<tr>
<td>Glutamate synthase</td>
<td>1ea0</td>
<td>65</td>
<td>72</td>
<td>N/A</td>
<td>-24.87</td>
<td>N/A</td>
</tr>
<tr>
<td>Pectin methylesterase</td>
<td>1qv</td>
<td>11</td>
<td>146</td>
<td>266</td>
<td>-20.74</td>
<td>1</td>
</tr>
<tr>
<td>P22 tailspike</td>
<td>1tyu</td>
<td>30</td>
<td>15</td>
<td>2</td>
<td>-20.46</td>
<td>1</td>
</tr>
<tr>
<td>Iota-carrageenase</td>
<td>1ktw</td>
<td>17</td>
<td>121</td>
<td>10</td>
<td>-23.4</td>
<td>N/A</td>
</tr>
<tr>
<td>Pectate lyase</td>
<td>1air</td>
<td>2</td>
<td>35</td>
<td>45</td>
<td>-16.02</td>
<td>1</td>
</tr>
<tr>
<td>1bn8</td>
<td>1</td>
<td>37</td>
<td>76</td>
<td>-18.42</td>
<td>3</td>
<td>12</td>
</tr>
<tr>
<td>1ee6</td>
<td>852</td>
<td>880</td>
<td>228</td>
<td>-16.44</td>
<td>2</td>
<td>10.84</td>
</tr>
<tr>
<td>Pectin lyase</td>
<td>1idj</td>
<td>14</td>
<td>257</td>
<td>6</td>
<td>-17.99</td>
<td>2</td>
</tr>
<tr>
<td>1qcx</td>
<td>28</td>
<td>263</td>
<td>6</td>
<td>-17.09</td>
<td>1</td>
<td>16.43</td>
</tr>
<tr>
<td>Galacturonase</td>
<td>1bhe</td>
<td>18</td>
<td>108</td>
<td>18</td>
<td>-18.80</td>
<td>1</td>
</tr>
<tr>
<td>1czf</td>
<td>43</td>
<td>130</td>
<td>5</td>
<td>-19.32</td>
<td>2</td>
<td>40.37</td>
</tr>
<tr>
<td>1rimg</td>
<td>3</td>
<td>270</td>
<td>27</td>
<td>-20.12</td>
<td>3</td>
<td>23.93</td>
</tr>
</tbody>
</table>
Structural Motif Recognition

Experiment Results on SCRF for $\beta$-helix

- Predicted Segmentation for known $\beta$-helix

<table>
<thead>
<tr>
<th>Group</th>
<th>Perfect match</th>
<th>Good match</th>
<th>OK match</th>
</tr>
</thead>
<tbody>
<tr>
<td>Missing rungs</td>
<td>0</td>
<td>1-2</td>
<td>3 or more</td>
</tr>
<tr>
<td>PDB-ID</td>
<td>1czf</td>
<td>1air, 1bhe, 1bn8, 1dbc, 1ee6(right), 1idj, 1ktw(left), 1qex, 1qjv, 1rmg</td>
<td>1dab(left), 1ea0, 1tyu(right)</td>
</tr>
</tbody>
</table>
Structural Motif Recognition

Experiment Results on SCRF for $\beta$-helix

- Histograms of scores for known $\beta$-helices against PDB-minus dataset on cross-family validation

![Histogram of scores for known $\beta$-helices against PDB-minus dataset on cross-family validation.](image)
Verification on Recently Crystallized Structures

Successfully identify proteins from different organisms

- 1YP2: Potato Tuber ADP-Glucose Pyrophosphorylase
  - score 10.47

- 1PXZ: Jun A 1, The Major Allergen From Cedar Pollen
  - score 32.35

- GP14 of Shigella bacteriophage as a β-helix protein with scoring 15.63
Experiment Results on Chain Graph Model for β-helix

- Cross-family validation for classifying β-helix proteins
  - Chain graph model can score all known β-helices higher than non β-helices

<table>
<thead>
<tr>
<th>SCOP Family</th>
<th>PDB-ID</th>
<th>Struct-based HMMs</th>
<th>Threader</th>
<th>BetaWrap1</th>
<th>SCRPs</th>
<th>CGM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Bit score Rank</td>
<td></td>
<td>Wrap-score Rank</td>
<td></td>
<td></td>
</tr>
<tr>
<td>P.69 pertactin</td>
<td>1DAB</td>
<td>-73.6 3</td>
<td>24</td>
<td>-17.84 1</td>
<td>10.17 1</td>
<td>31.69 1</td>
</tr>
<tr>
<td>Chondroitinase B</td>
<td>1DBG</td>
<td>-64.6 5</td>
<td>47</td>
<td>-19.55 1</td>
<td>13.15 1</td>
<td>34.89 1</td>
</tr>
<tr>
<td>Glutamate synthase</td>
<td>1EA0</td>
<td>-85.7 65</td>
<td>N/A</td>
<td>-24.87 N/A</td>
<td>6.21 1</td>
<td>29.04 1</td>
</tr>
<tr>
<td>Pectin methylesterase</td>
<td>1QJY</td>
<td>-72.8 11</td>
<td>266</td>
<td>-20.74 1</td>
<td>6.12 1</td>
<td>22.69 1</td>
</tr>
<tr>
<td>P22 tailspike</td>
<td>1TYU</td>
<td>-78.8 30</td>
<td>2</td>
<td>-20.46 1</td>
<td>6.71 1</td>
<td>20.59 1</td>
</tr>
<tr>
<td>Iota-carrageeenase</td>
<td>1KTW</td>
<td>-81.9 17</td>
<td>10</td>
<td>-23.4 N/A</td>
<td>8.07 1</td>
<td>16.06 1</td>
</tr>
<tr>
<td>Pectate lyase</td>
<td>1AIR</td>
<td>-37.1 2</td>
<td>45</td>
<td>-16.02 1</td>
<td>16.64 1</td>
<td>22.87 2</td>
</tr>
<tr>
<td></td>
<td>1BN8</td>
<td>180.3 1</td>
<td>76</td>
<td>-18.42 3</td>
<td>13.28 2</td>
<td>28.98 1</td>
</tr>
<tr>
<td></td>
<td>1EE6</td>
<td>-170.8 852</td>
<td>228</td>
<td>-16.44 2</td>
<td>10.84 3</td>
<td>15.16 3</td>
</tr>
<tr>
<td>Pectin lyase</td>
<td>1Dj</td>
<td>-78.1 14</td>
<td>6</td>
<td>-17.99 2</td>
<td>15.01 2</td>
<td>17.50 2</td>
</tr>
<tr>
<td></td>
<td>1QCX</td>
<td>-83.5 28</td>
<td>6</td>
<td>-17.09 1</td>
<td>16.43 1</td>
<td>20.67 1</td>
</tr>
<tr>
<td>Galacturonase</td>
<td>1BHE</td>
<td>-91.5 18</td>
<td>18</td>
<td>-18.80 1</td>
<td>20.11 3</td>
<td>28.98 1</td>
</tr>
<tr>
<td></td>
<td>1CZF</td>
<td>-98.4 43</td>
<td>5</td>
<td>-19.32 2</td>
<td>40.37 1</td>
<td>24.68 3</td>
</tr>
<tr>
<td></td>
<td>1RMG</td>
<td>-76.3 3</td>
<td>27</td>
<td>-20.12 3</td>
<td>23.93 2</td>
<td>27.37 2</td>
</tr>
</tbody>
</table>

Chain graph model reduces the real running time by around 50 times faster
Experiment Results on Chain Graph Model for LLR

- Cross-family validation for classifying LLR proteins
  - Chain graph model can score all known LLR higher than non-LLR

<table>
<thead>
<tr>
<th>SCOP Family</th>
<th>PDB-ID</th>
<th>ClustalW+HMMs</th>
<th>Struct-based HMMs</th>
<th>Threader</th>
<th>CGM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Bit score</td>
<td>Rank</td>
<td>Bit Score</td>
<td>Rank</td>
</tr>
<tr>
<td>28-residue LRR</td>
<td>1A4Y</td>
<td>-123.5</td>
<td>4</td>
<td>-76.7</td>
<td>1</td>
</tr>
<tr>
<td>Rna1p (RanGAP1)</td>
<td>1YRG</td>
<td>-95.4</td>
<td>1</td>
<td>-81.1</td>
<td>1</td>
</tr>
<tr>
<td>Cyclin A/CDK2-associated p19</td>
<td>1FQV</td>
<td>-163.3</td>
<td>89</td>
<td>-111.4</td>
<td>10</td>
</tr>
<tr>
<td>Internalin LRR domain</td>
<td>1O6V</td>
<td>-52.8</td>
<td>1</td>
<td>-0.7</td>
<td>1</td>
</tr>
<tr>
<td>Leucine rich effector</td>
<td>1JL5</td>
<td>-86.7</td>
<td>1</td>
<td>-26.5</td>
<td>1</td>
</tr>
<tr>
<td>Ngr ectodomain-like</td>
<td>1P9A</td>
<td>-120.0</td>
<td>9</td>
<td>-68.6</td>
<td>1</td>
</tr>
<tr>
<td>Polygalacturonase inhibiting protein</td>
<td>1OGQ</td>
<td>-155.0</td>
<td>32</td>
<td>-18.2</td>
<td>1</td>
</tr>
<tr>
<td>Rab geranylgeranyltransferase alpha-subunit</td>
<td>1DCE</td>
<td>-145.4</td>
<td>16</td>
<td>-59.7</td>
<td>1</td>
</tr>
<tr>
<td>mRNA export factor</td>
<td>1KOH</td>
<td>-153.9</td>
<td>42</td>
<td>-91.7</td>
<td>1</td>
</tr>
<tr>
<td>U2A-like</td>
<td>1A9N</td>
<td>-280.9</td>
<td>861</td>
<td>-151.4</td>
<td>478</td>
</tr>
<tr>
<td>L domain</td>
<td>1IGR</td>
<td>-150.0</td>
<td>46</td>
<td>-107.1</td>
<td>249</td>
</tr>
</tbody>
</table>
Experiment Results on Chain Graph Model

- Predicted Segmentation for known β-helices and LLRs
  - (A) SCRFs; (B) chain graph model
Further Experiments

- Virus proteins
  - Noncellular biological entity that can reproduce only within a host cell
  - Dynamical properties
  - Successes in virus-related proteins
    - Adenovirus (cause for common cold)
    - Bacteriophage (which infects bacteria)
    - DNA virus, RNA virus

Gp 41: core protein of HIV virus (1AIK)
Summary

- Segmented conditional graphical models for structural motif recognition
  - Segmentation conditional random fields
  - Chain graph model

- Successful applications with limited positive examples 😊😊😊
  - Right-handed beta-helices
  - Leucine-rich repeats

- Further verification
  - Virus spike folds and others
## Outline

Conditional graphical models for protein structure prediction

<table>
<thead>
<tr>
<th>Hiarchy</th>
<th>Secondary Structures</th>
<th>Tertiary Structures</th>
<th>Quaternary Structures</th>
</tr>
</thead>
<tbody>
<tr>
<td>Examples</td>
<td><img src="image1.png" alt="Example 1" /></td>
<td><img src="image2.png" alt="Example 2" /></td>
<td><img src="image3.png" alt="Example 3" /></td>
</tr>
<tr>
<td>Structural building blocks</td>
<td>Amino acids (residues)</td>
<td>Secondary structure elements</td>
<td>Secondary structure and tertiary structure elements</td>
</tr>
<tr>
<td>Specific model</td>
<td>CRFs Kernel CRFs</td>
<td>Segmentation CRFs Chain graph model</td>
<td>Factorial Segmentation CRFs</td>
</tr>
<tr>
<td>Status</td>
<td>Completed</td>
<td>Work in progress</td>
<td>Proposed work</td>
</tr>
</tbody>
</table>
Quaternary Structure Prediction

- Quaternary structures
  - Multiple chains associated together through noncovalent bonds or disulfide bonds

- Classes of quaternary structures
  - Based on the number of subunits
    - Usually 2 - 10
  - Based on the identity of the subunits
    - Homo-oligomers* and hetero-oligomers

- Very limited research work to date

..APAFSVSPASGACGPECA..
Contains the quaternary structures?
Yes

..NNEEEECCCCCHHHCCC..
Proposed Work

- **Structural component**
  - Secondary structure elements
  - Super-secondary structure elements

- **Protein structural graph**
  - Nodes for the states of secondary structural or super-secondary structural elements of unknown lengths
  - Edges for interactions between nodes in 3-D (inter- and intra-chain)
  - Protocol: nodes representing secondary structure elements must involve long-range interactions
Proposed Work

• Segmentation definition
  \[ W = (M, \{W_i\}), \]
  \( M: \# \text{ of segments} \)
  \( W_i: \text{configuration of segment } i \)

• Factorial segmentation conditional random fields
  \[ P(W^{(1)}, ..., W^{(R)} | x) = \frac{1}{Z} \prod_{c \in C_g} \exp(\sum_{k=1}^{K} \lambda_k f_k(x, w_c)) \]
  \( R: \# \text{ of chains in quaternary structures} \)
  \( W^{(1)}_i, W^{(2)}_i, ..., W^{(R)}_i \) have different interaction maps
Experiment Design

- **Triple beta-spirals**
  - Described by van Raaij et al. in Nature (1999)
  - Clear sequence repeats
  - Two proteins with crystallized structures and about 20 without structure annotation

- **Tripe beta-helices**
  - Described by van Raaij et al. in JMB (2001)
  - Structurally similar to beta-helix without clear sequence similarity
  - Two proteins with crystallized structures
  - Information from beta-helices can be used

- Both folds characterized by unusual stability to heat, protease, and detergent
## Summary

### Conditional graphical models for protein structure prediction

<table>
<thead>
<tr>
<th>Secondary Structures</th>
<th>Tertiary Structures</th>
<th>Quaternary Structures</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image" alt="Amino acids (residues)" /></td>
<td>Secondary structure elements</td>
<td>Secondary structure and tertiary structure elements</td>
</tr>
<tr>
<td>CRFs&lt;br&gt;Kernel CRFs</td>
<td>Segmentation CRFs&lt;br&gt;Chain graph model</td>
<td>Factorial Segmentation CRFs</td>
</tr>
<tr>
<td>Globular proteins</td>
<td>Beta-helix, Leucine-rich repeats and other folds</td>
<td>Triple beta-spirals, triple beta-helices and other folds</td>
</tr>
<tr>
<td>Completed</td>
<td>Work in progress</td>
<td>Work in progress</td>
</tr>
</tbody>
</table>
Expected Contribution

• Computational contribution
  ➢ Conditional graphical models for structured data prediction
  ➢ “First effective models for the long-range interactions”

• Biological contribution
  ➢ Improvement in protein structure prediction
  ➢ Hypothesis on potential proteins with biological important folds
  ➢ Aids for function analysis and drug design
Timeline

- Jun, 2005 -- Aug, 2005
  - Data collection for triple beta-spirals and triple beta-helices
  - Preliminary investigation for virus protein folds

- Sep, 2005 -- Nov, 2005
  - Implementation and Testing the model on synthetic data and real data
  - Determining the specific virus proteins to work on

- Nov, 2005 -- Jan, 2006
  - Virus protein fold recognition
  - Analysis of the properties for virus protein folds

- Feb, 2006 -- June, 2006:
  - Virus protein fold recognition
  - Investigation the possibility for protein function prediction or information extraction

  - Writing up thesis
Features

• Node features
  - Regular expression template, HMM profiles
  - Secondary structure prediction scores
  - Segment length

• Inter-node features
  - $\beta$-strand Side-chain alignment scores
  - Preferences for parallel alignment scores
  - Distance between adjacent B23 segments

• Features are general and easy to extend
Evaluation Measure

- **Q3 (accuracy)**
  \[ Q_3 = \frac{p + n}{p + n + o + u} \]

- **Precision, Recall**
  \[ Q^{pre} = \frac{p}{p + o} \quad Q = \frac{p}{p + u} \]

- **Segment Overlap quantity (SOV)**
  \[ SOV(S1, S2) = \frac{1}{N} \sum_{S(i)} \frac{MINOV(S1; S2) + \text{DELTA}(S1; S2)}{\text{MAXOV}(S1; S2)} \cdot \text{LEN}(S1) \]

- **Matthew’s Correlation coefficients**
  \[ C_i = \frac{p_i \cdot n_i - u_i \cdot o_i}{\sqrt{(p_i + u_i) \cdot (p_i + o_i) \cdot (n_i + u_i) \cdot (n_i + o_i)}} \]
Local Information

• PSI-blast profile
  - Position-specific scoring matrices (PSSM)
  - Linear transformation [Kim & Park, 2003]

\[ f(x) = \begin{cases} 
  0 & \text{if } (x \leq -5) \\
  0.5 + 0.1x & \text{if } (-5 < x < 5) \\
  1.0 & \text{if } (x \geq 5) 
\end{cases} \]

• SVM classifier with RBF kernel
• Feature#1 (Si): Prediction score for each residue Ri
Previous Work

• Huge literature over decades
  ➢ Window-based methods
  ➢ Hidden Markov models

• Major breakthroughs in recent years
  ➢ Combine the predictions from neighboring residues or various methods (Cuff & Barton, 1999)
  ➢ Explore evolutionary information from sequences (Jones, 1999)
  ➢ Specific algorithm for beta-sheets or infer the paring of beta-strands (Meiler & Baker, 2003)