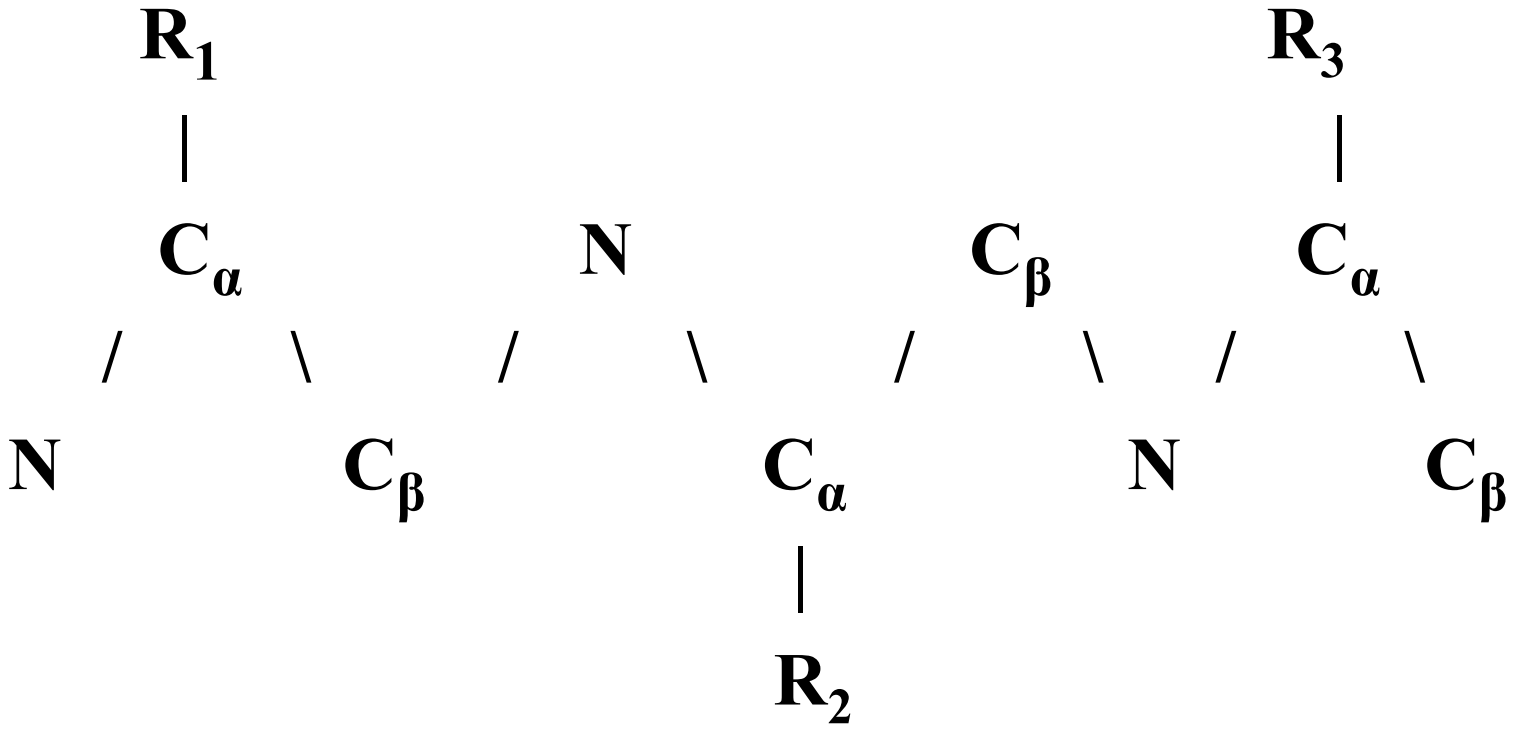
# Bioinformatics and Machine Learning: the Prediction of Protein Structures on a Genomic Scale

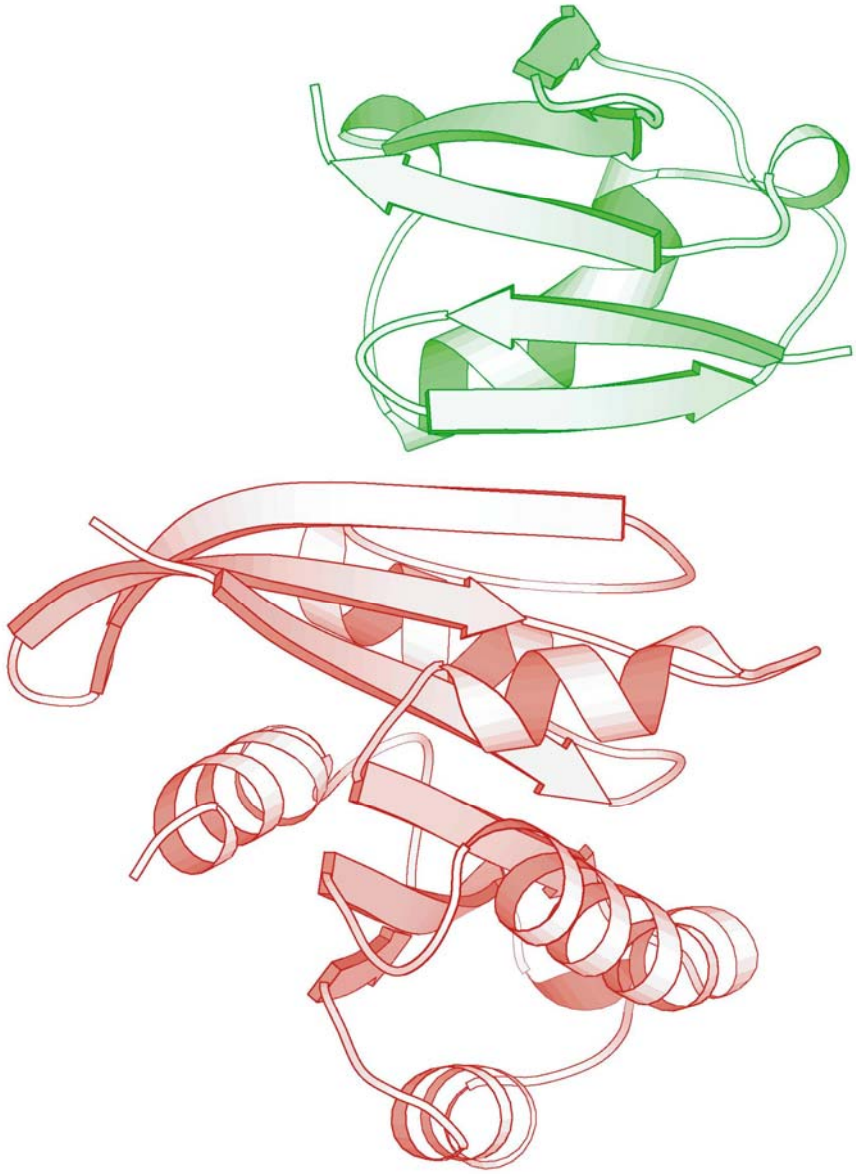
Pierre Baldi
Dept. Information and Computer Science
Institute for Genomics and Bioinformatics
University of California, Irvine

# UNDERSTANDING INTELLIGENCE

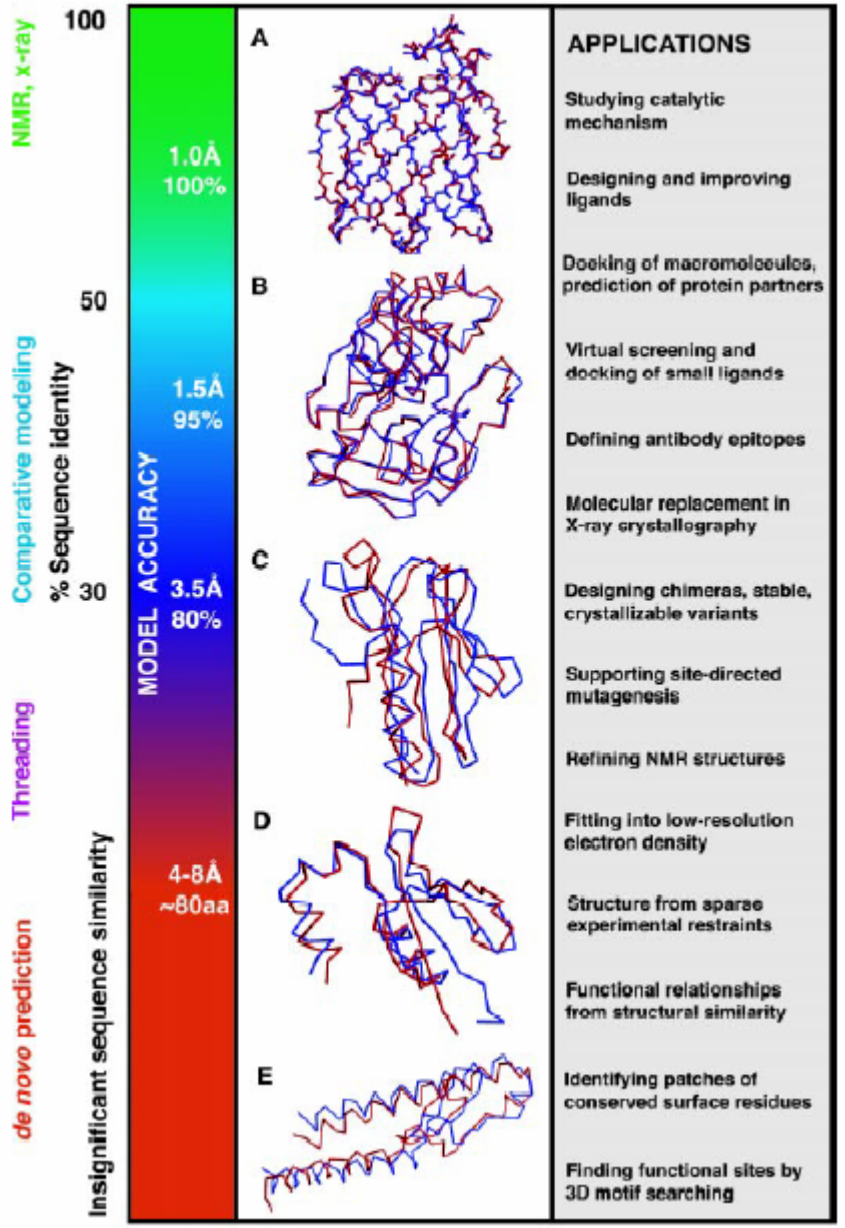
- **Human intelligence (inverse problem)**
- **AI (direct problem)**
- **Intelligence?**
- **Solve specific problems.**
- **Choice of problems is key.**

# PROTEINS



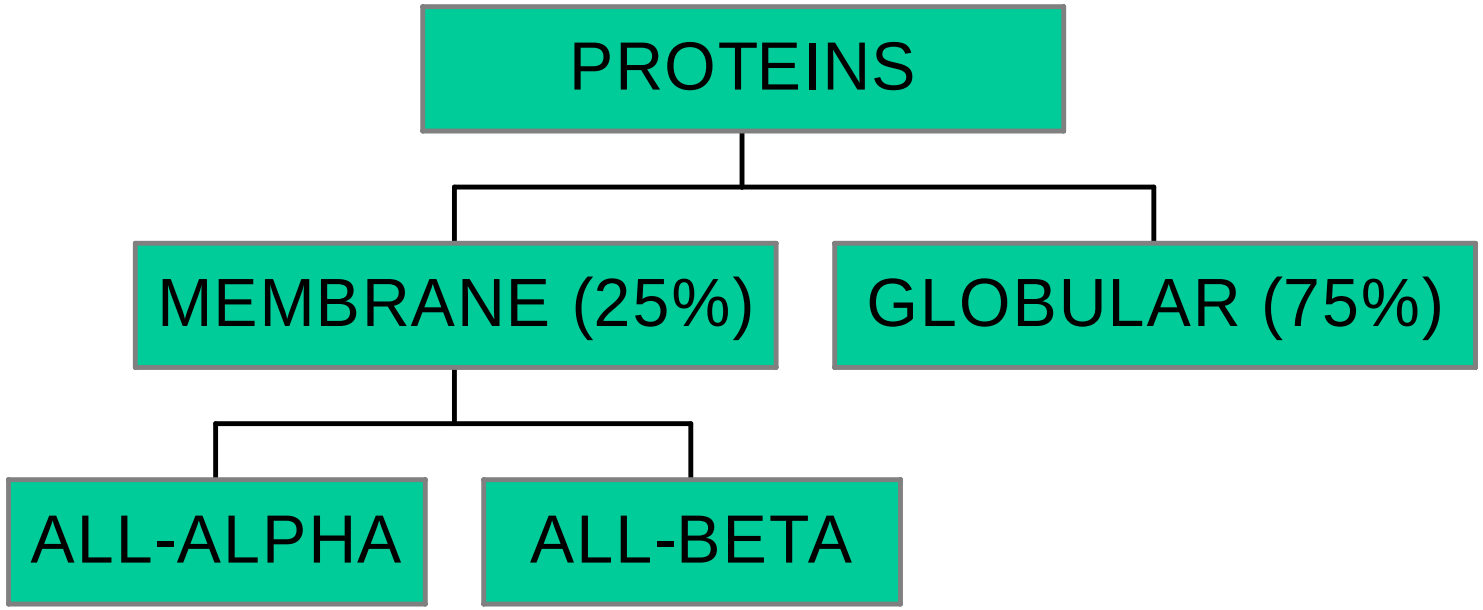


# Utility of Structural Information




(Baker and Sali, 2001)

# CAVEAT



# REMARKS

- **Structure/Folding**
- **Backbone/Full Atom**
- **Homology Modeling**
- **Protein Threading**
- **Ab Initio (Physical Potentials/Molecular Dynamics, Statistical Mechanics/Lattice Models)**
- **Statistical/Machine Learning (Training Sets, SS prediction)**
- **Mixtures: ab-initio with statistical potentials, machine learning with profiles, etc.**

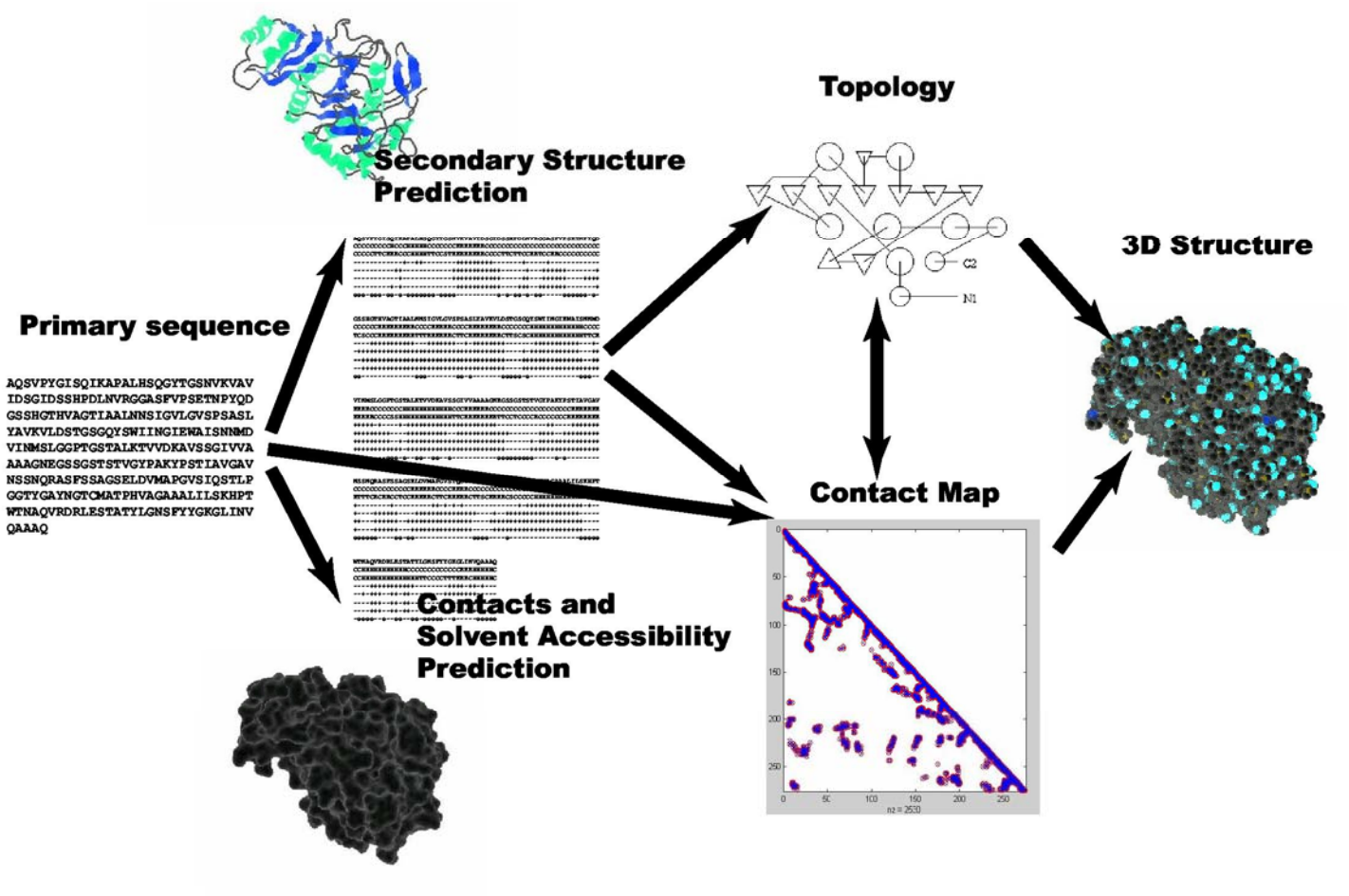
# PROTEIN STRUCTURE PREDICTION

**DECOMPOSITION INTO 3 PROBLEMS**

1. **FROM PRIMARY SEQUENCE TO SECONDARY STRUCTURE AND OTHER STRUCTURAL FEATURES**

2. **FROM PRIMARY SEQUENCE AND STRUCTURAL FEATURES TO TOPOLOGICAL REPRESENTATION**

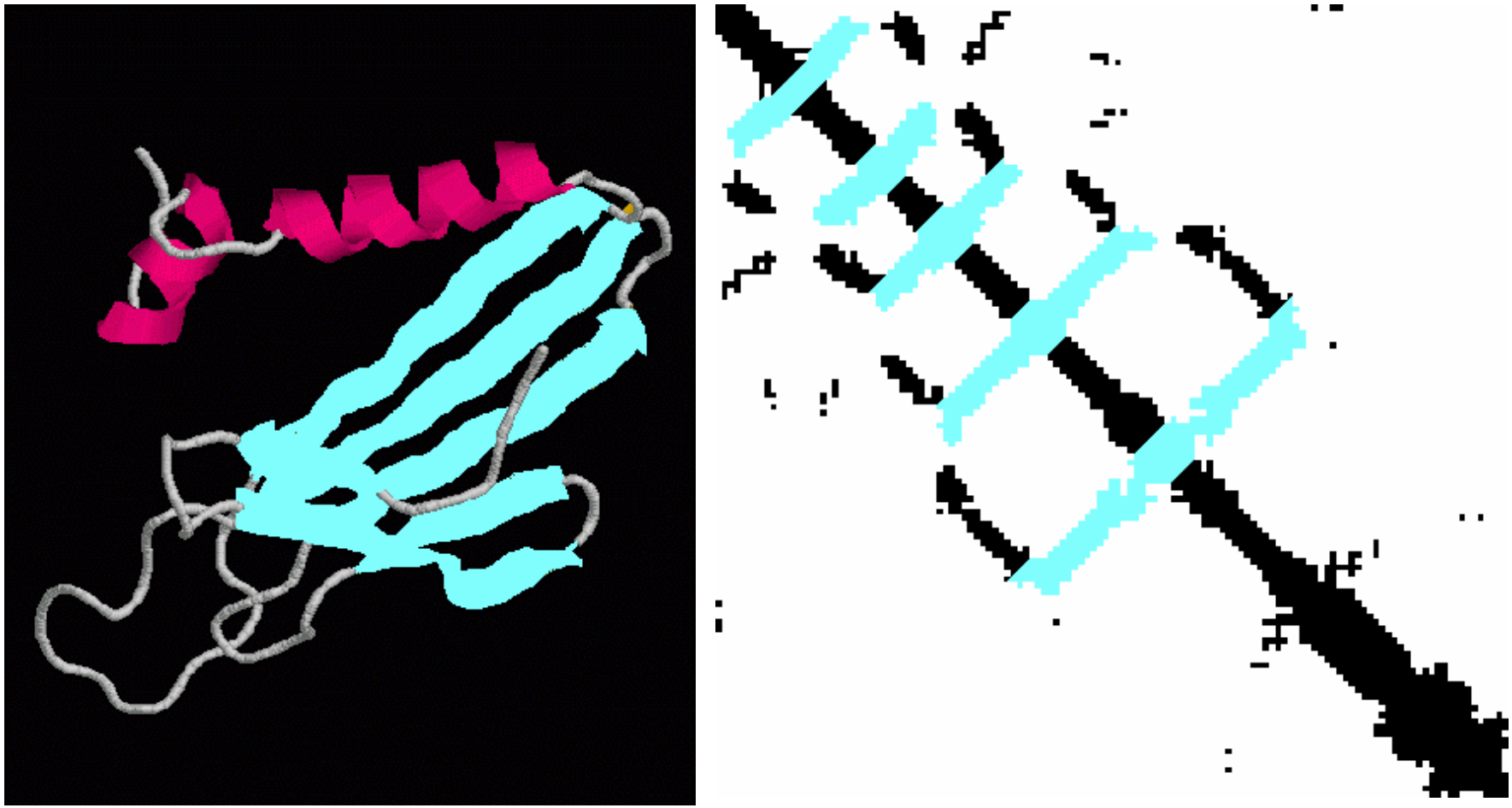
3. **FROM TOPOLOGICAL REPRESENTATION TO 3D COORDINATES**


Secondary Structure Prediction
Topology
3D Structure
Primary sequence
AQSVPYGISQIKAPALHSQGYTGSNVKVAV
IDSGIDSSHPDLNVRGGASFVPSETNPYQD
GSSHGTHVAGTIAALNNSIGVLGVSPSASL
YAVKVLDSTGSGQYSWIINGIEWAISNNMD
VINMSLGGPTGSTALKTVVDKAVSSGIVVA
AAAGNEGSSGSTSTVGYPAKYPSTIAVGAV
NSSNQRASFSSAGSELDVMAPGVSIQSTLP
GGTYGAYNGTCMATPHVAGAAALILSKHPT
WTNAQVRDRLESTATYLGNSFYYGKGLINV
QAAAQ
Contact Map
Contacts and Solvent Accessibility Prediction

# Helices

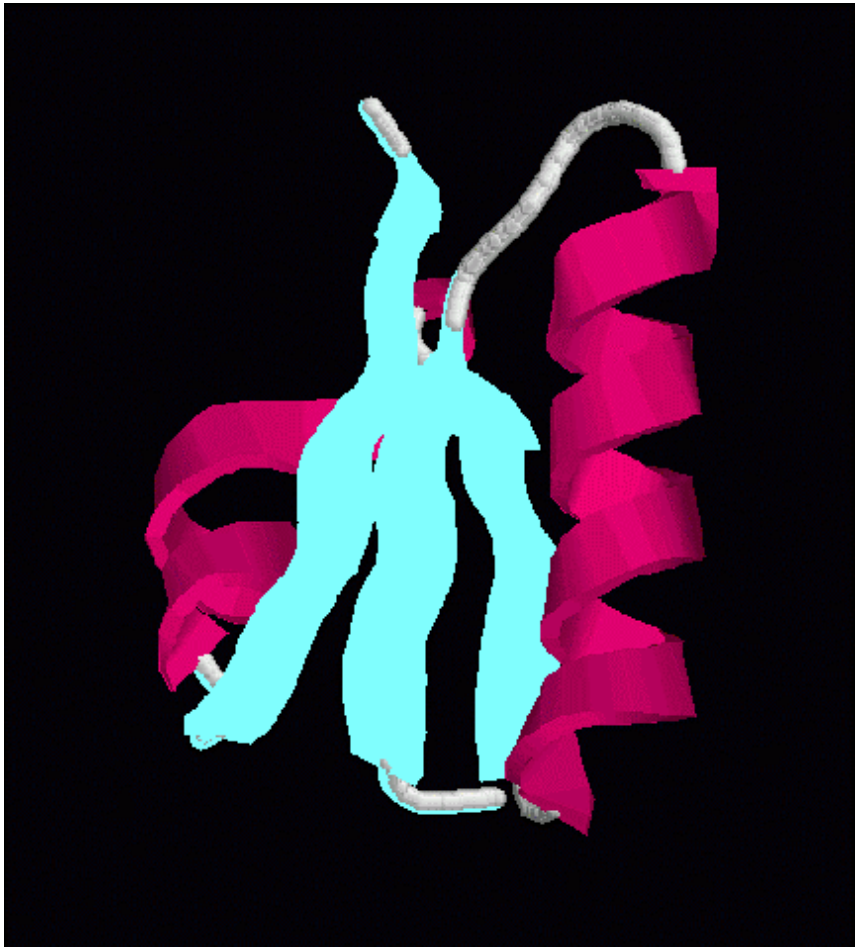
**1GRJ (Grea Transcript Cleavage Factor From Escherichia Coli)**

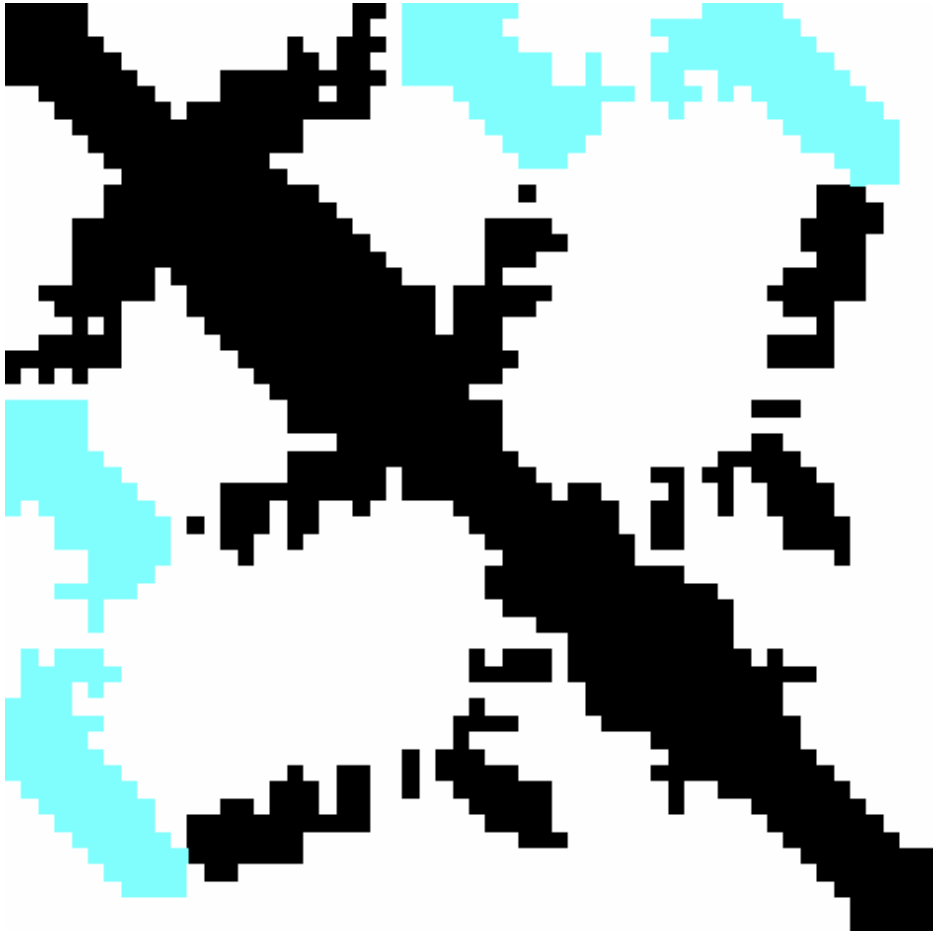
# Antiparallel β-sheets



**1MSC (Bacteriophage Ms2 Unassembled Coat Protein Dimer)**

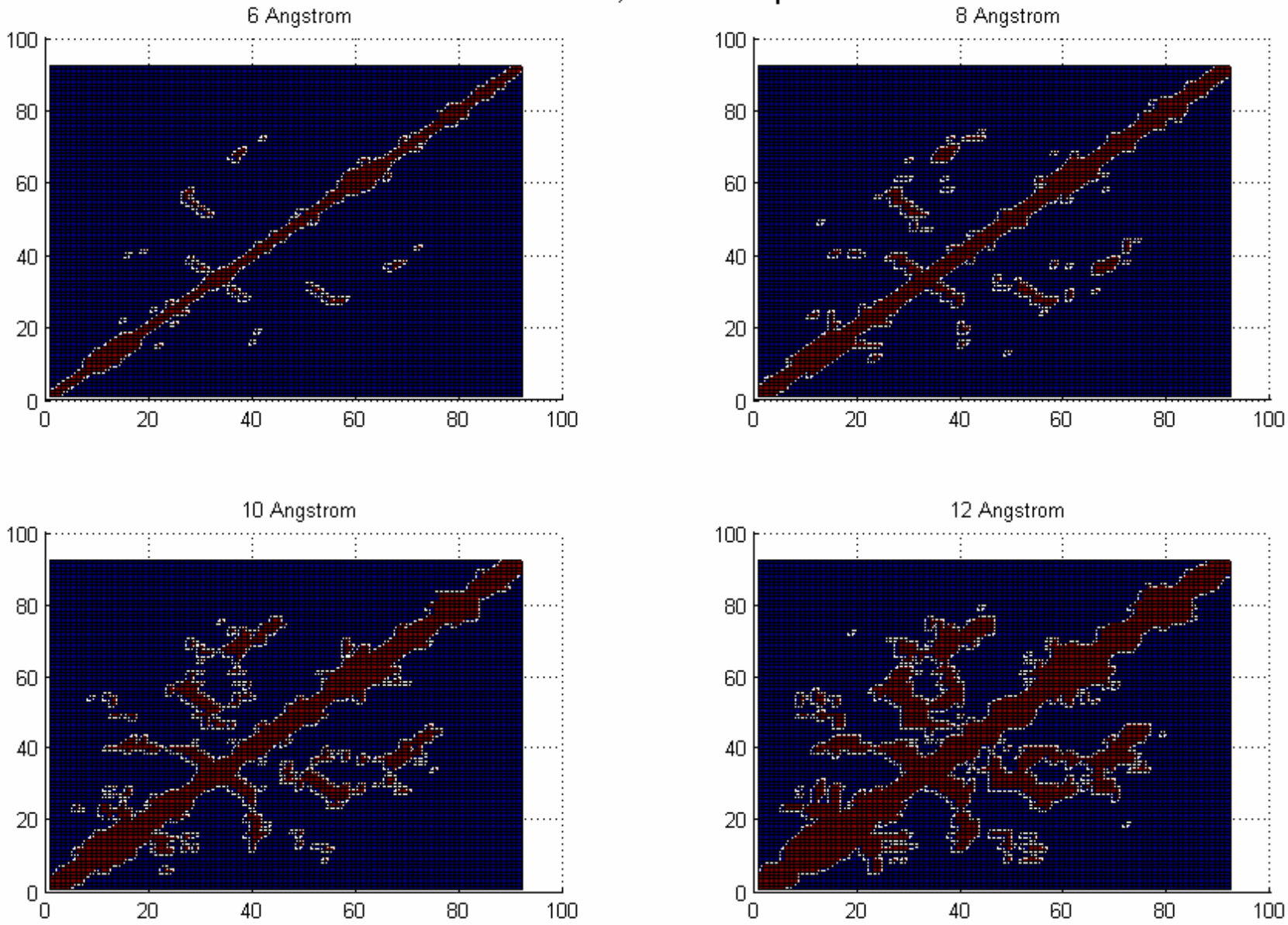
# Parallel β-sheets



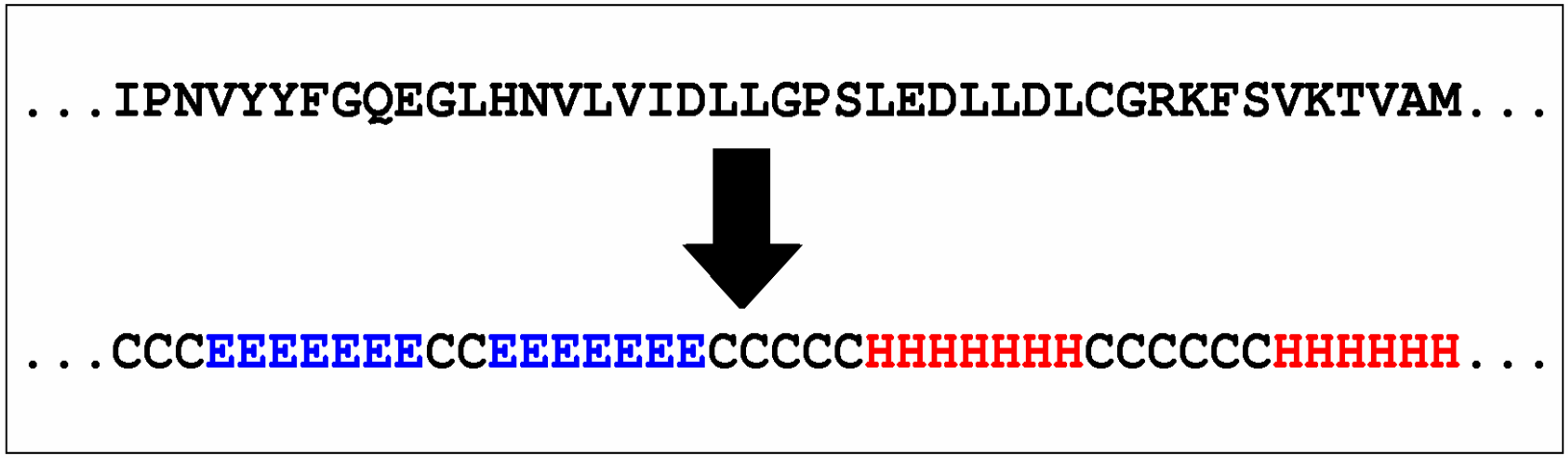


**1FUE (Flavodoxin)**

# 1ADNA, contact maps



# Secondary structure prediction

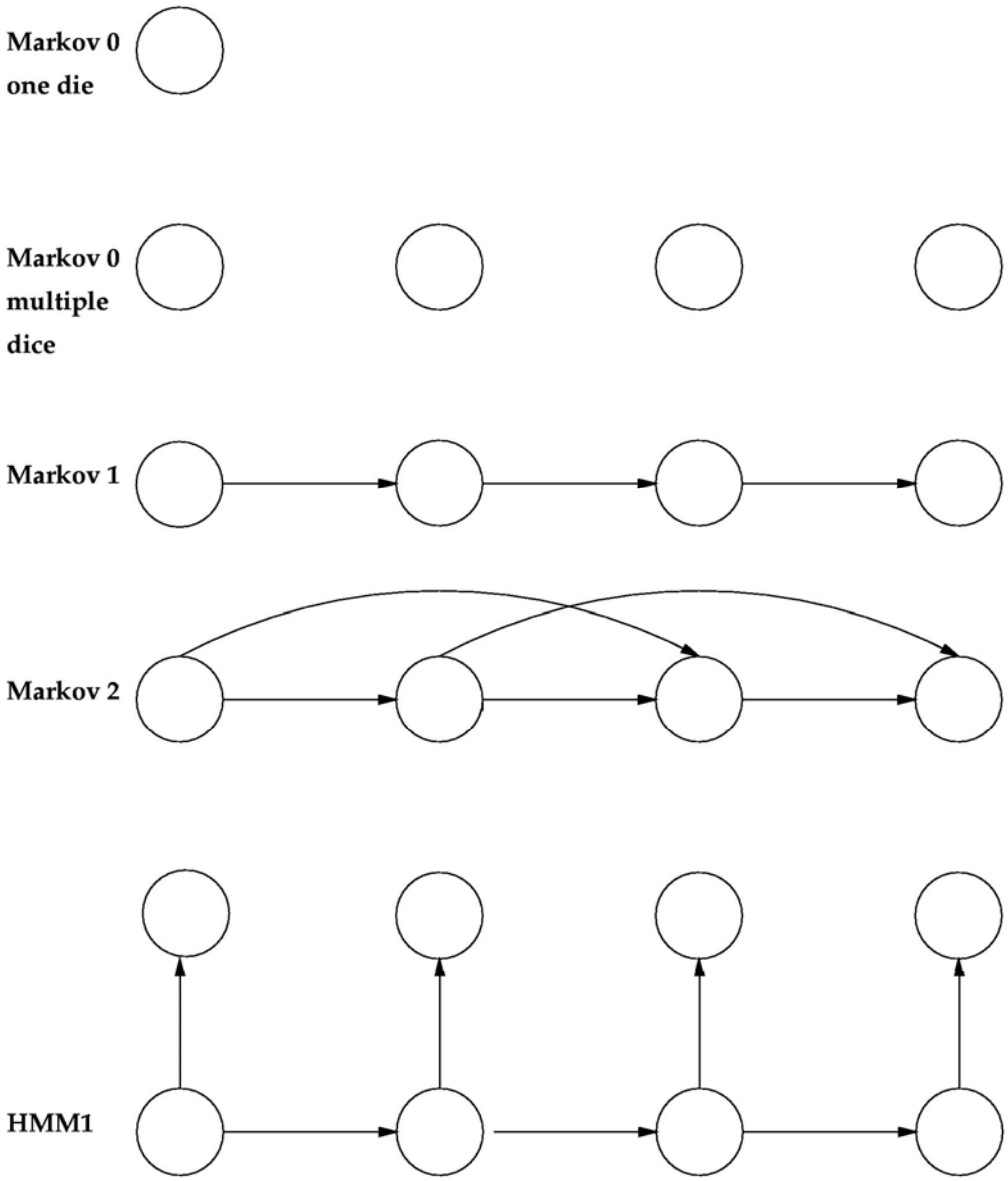


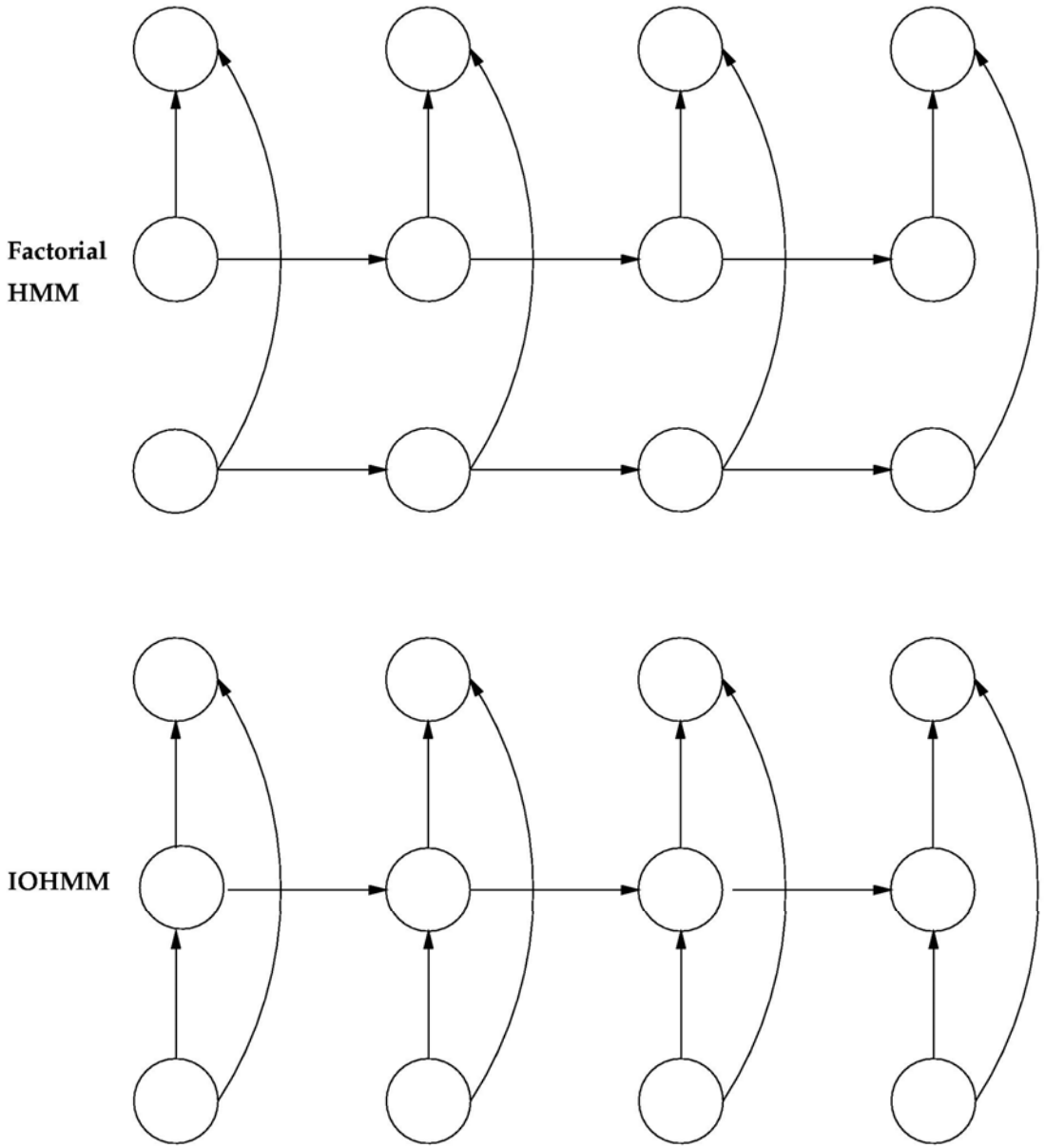
# GRAPHICAL MODELS: BAYESIAN NETWORKS

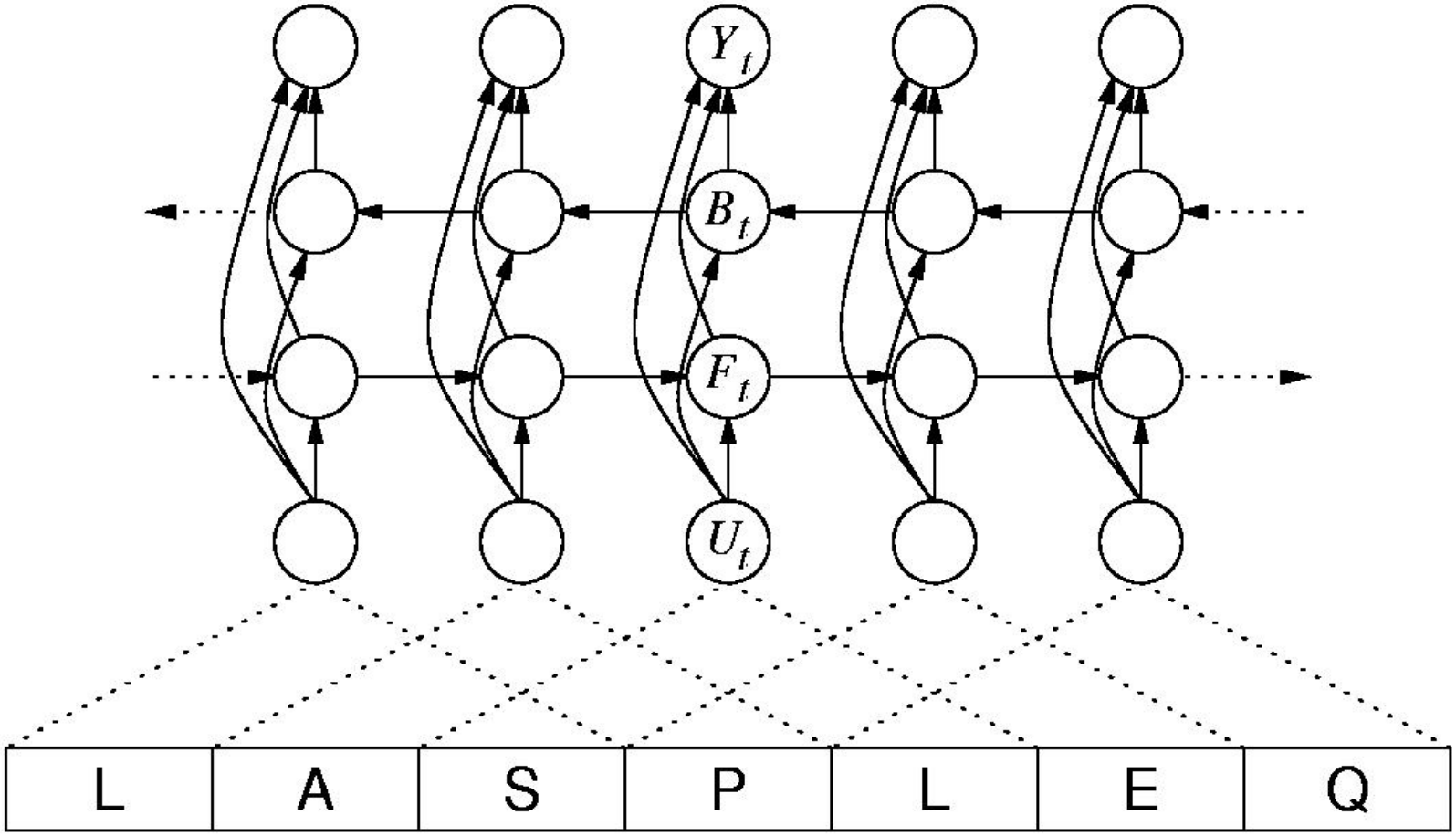
- **X1, … ,Xn random variables associated with the vertices of a DAG = Directed Acyclic Graph**
- **The local conditional distributions P(Xi|Xj: j parent of i) are the parameters of the model. They can be represented by look-up tables (costly) or other more compact parameterizations (Sigmoidal Belief Networks, XOR, etc).**

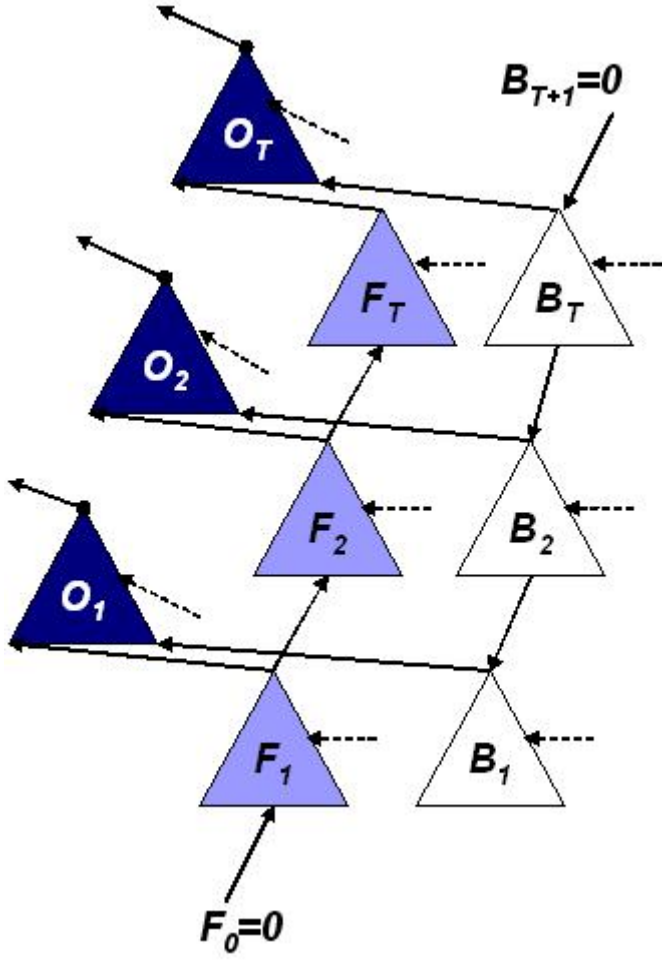
- **The global distribution is the product of the local characteristics:**

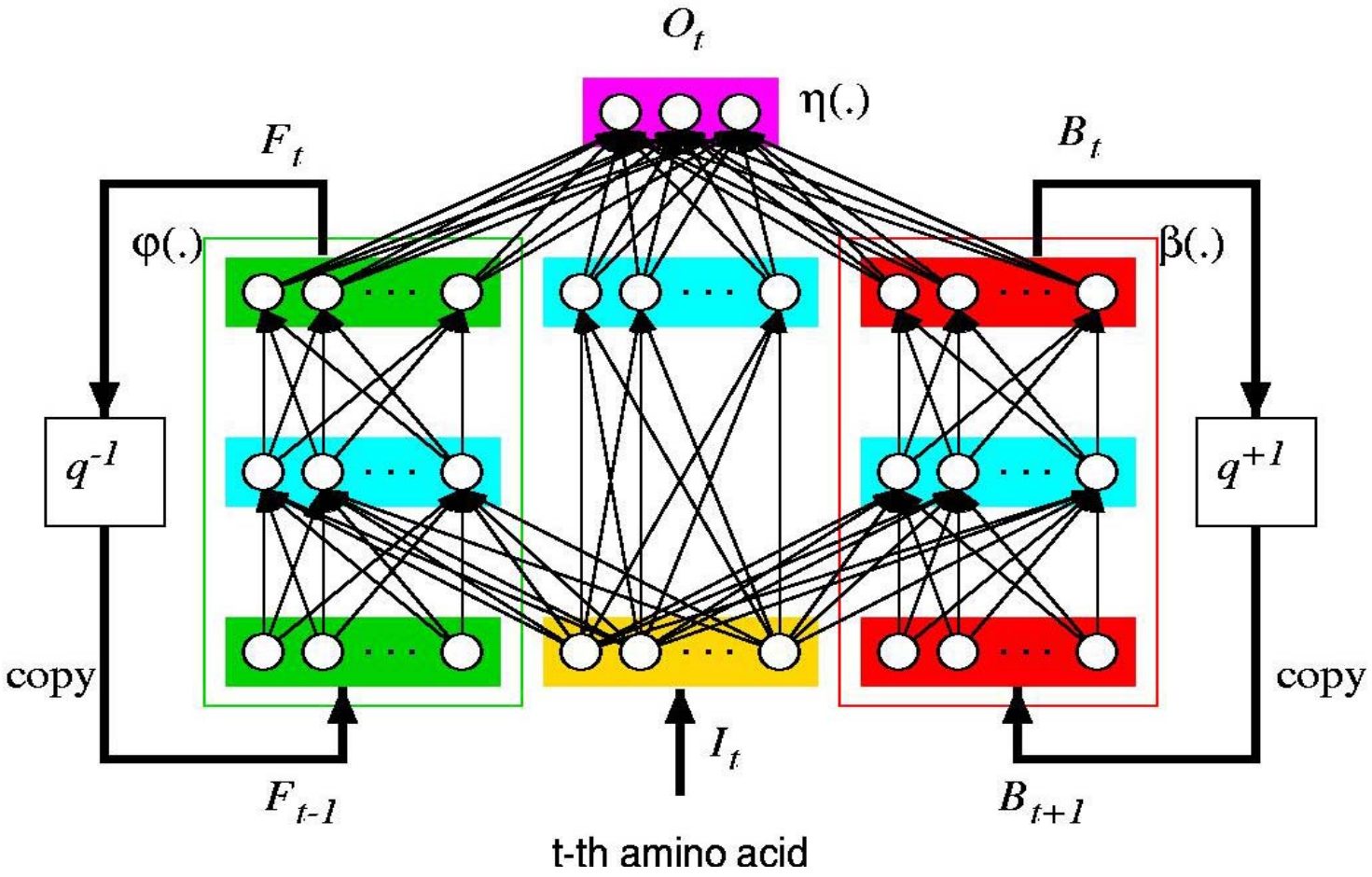
$$P(X1,\ldots,Xn) = \Pi_i P(Xi|Xj : j \text{ parent of } i)$$


Markov 0
one die
Markov 0
multiple
dice
Markov 1
Markov 2
HMM1


Factorial
HMM
IOHMM


$Y_t$
$B_t$
$F_t$
$U_t$
L
A
S
P
L
E
Q


$B_{T+1}=0$
$O_T$
$F_T$
$B_T$
$O_2$
$F_2$
$B_2$
$O_1$
$F_1$
$B_1$
$F_0=0$


$O_t$
η(.)
$F_t$
$B_t$
φ(.)
β(.)
$q^{-1}$
$q^{+1}$
copy
copy
$I_t$
$F_{t-1}$
$B_{t+1}$
t-th amino acid

# DATA PREPARATION

**Starting point: PDB data base.**

- **Remove sequences not determined by X ray diffraction.**
- **Remove sequences where DSSP crashes.**
- **Remove proteins with physical chain breaks (neighboring AA having distances exceeding 4 Angstroms)**
- **Remove sequences with resolution worst than 2.5 Angstroms.**
- **Remove chains with less than 30 AA.**
- **Remove redundancy (Hobohm's algorithm, Smith-Waterman, PAM 120, etc.)**
- **Build multiple alignments (BLAST, PSI-BLAST, etc.)**

# SECONDARY STRUCTURE PROGRAMS

- DSSP **(Kabsch and Sander, 1983): works by assigning potential backbone hydrogen bonds (based on the 3D coordinates of the backbone atoms) and subsequently by identifying repetitive bonding patterns.**
- STRIDE **(Frishman and Argos, 1995): in addition to hydrogen bonds, it uses also dihedral angles.**
- DEFINE **(Richards and Kundrot, 1988): uses difference distance matrices for evaluating the match of interatomic distances in the protein to those from idealized SS.**

# SECONDARY STRUCTURE ASSIGNMENTS

**DSSP classes:**

- **H = alpha helix**
- **E = sheet**
- **G = 3-10 helix**
- **S = kind of turn**
- **T = beta turn**
- **B = beta bridge**
- **I = pi-helix (very rare)**
- **C = the rest**

**CASP (harder) assignment:**

- **α = H and G**
- **β = E and B**
- **γ = the rest**

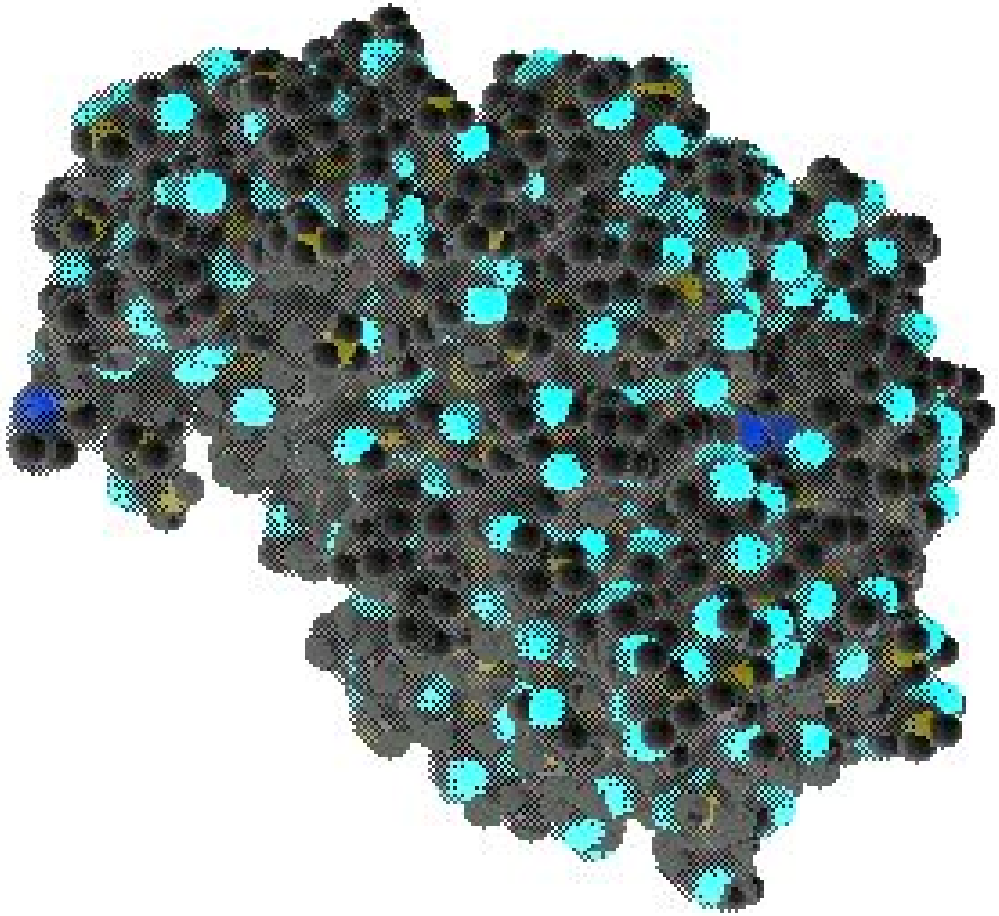
**Alternative assignment:**

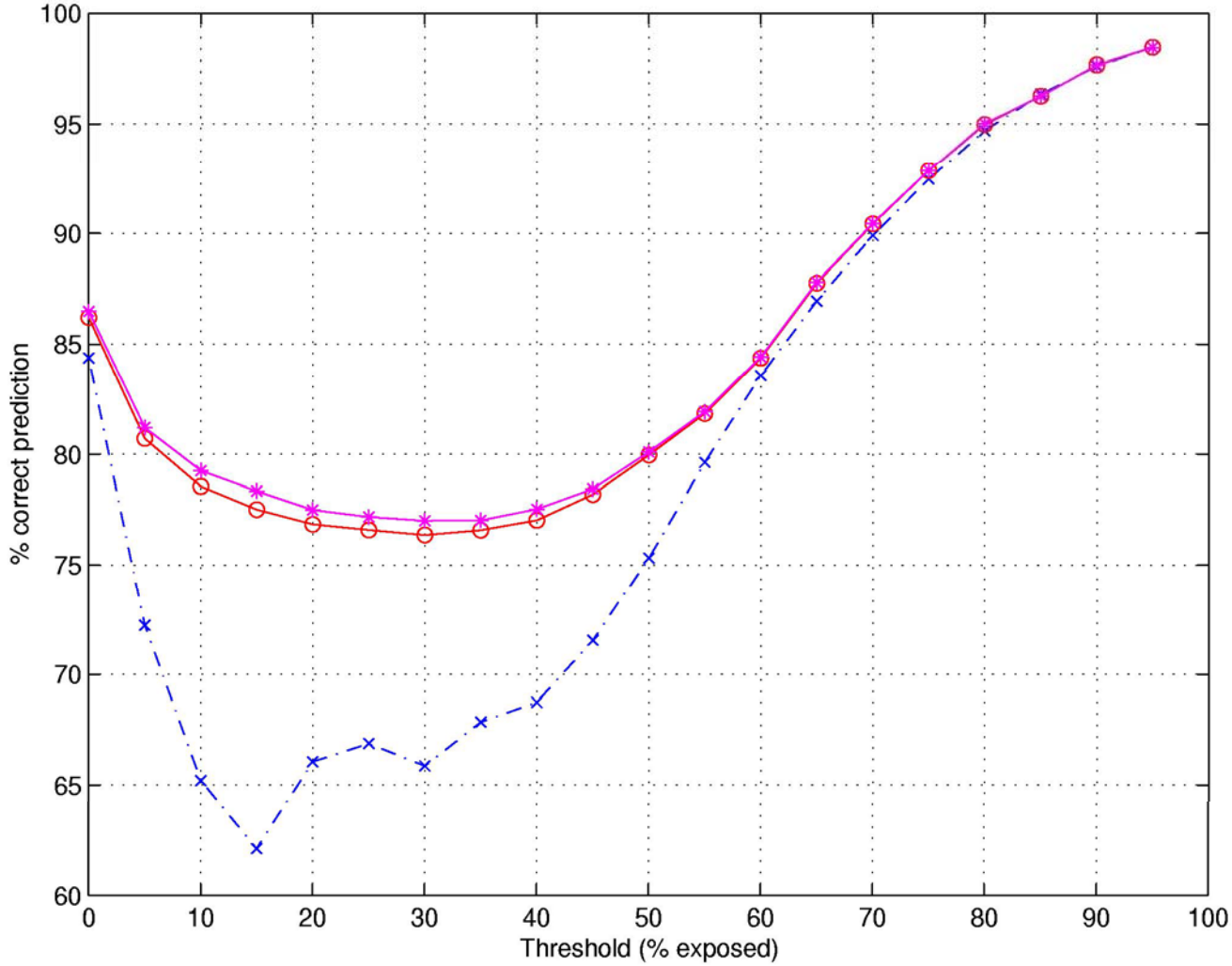
- **α = H**
- **β = B**
- **γ = the rest**

# ENSEMBLES

| Profiles | n | s | f | o | W | $Q_3$ residue |
|---|---|---|---|---|---|---|
| No | 7 | 2 | 8 | 11 | 1611 | 68.7% |
| No | 9 | 2 | 8 | 11 | 1899 | 68.8% |
| No | 7 | 3 | 8 | 11 | 1919 | 68.6% |
| No | 8 | 3 | 9 | 11 | 2181 | 68.8% |
| No | 20 | 0 | 17 | 11 | 2821 | 67.7% |
| Output | 9 | 2 | 8 | 11 | 1899 | 72.6% |
| Output | 8 | 3 | 9 | 11 | 2181 | 72.7% |
| Input | 9 | 2 | 8 | 11 | 1899 | 73.37% |
| Input | 8 | 3 | 9 | 11 | | 73.4% |
| Input | 12 | 3 | 9 | 10 | 2757 | 73.6% |
| Input | 7 | 3 | 8 | 11 | 1919 | 73.4% |
| Input | 8 | 3 | 9 | 10 | 2045 | 73.4% |
| Input | 12 | 3 | 9 | 11 | 2949 | 73.2% |

| SSpro 1.0 and SSpro 2.0 on the 3 test sets, Q3 | | | | |
|---|---|---|---|---|
| | | | | |
| | | SSpro 1.0 | SSpro 2.0 | |
| **R126** | H | 0.8079 | 0.8238 | |
| | E | 0.6323 | 0.6619 | |
| | C | 0.8056 | 0.8126 | |
| | **Q3** | **0.7662** | **0.7813** | |
| **EVA** | H | 0.8076 | 0.8248 | |
| | E | 0.625 | 0.6556 | |
| | C | 0.7805 | 0.7903 | |
| | **Q3** | **0.76** | **0.7767** | |
| **CASP4** | H | 0.8386 | 0.8608 | |
| | E | 0.6187 | 0.6851 | |
| | C | 0.8099 | 0.822 | |
| | **Q3** | **0.778** | **0.8065** | |

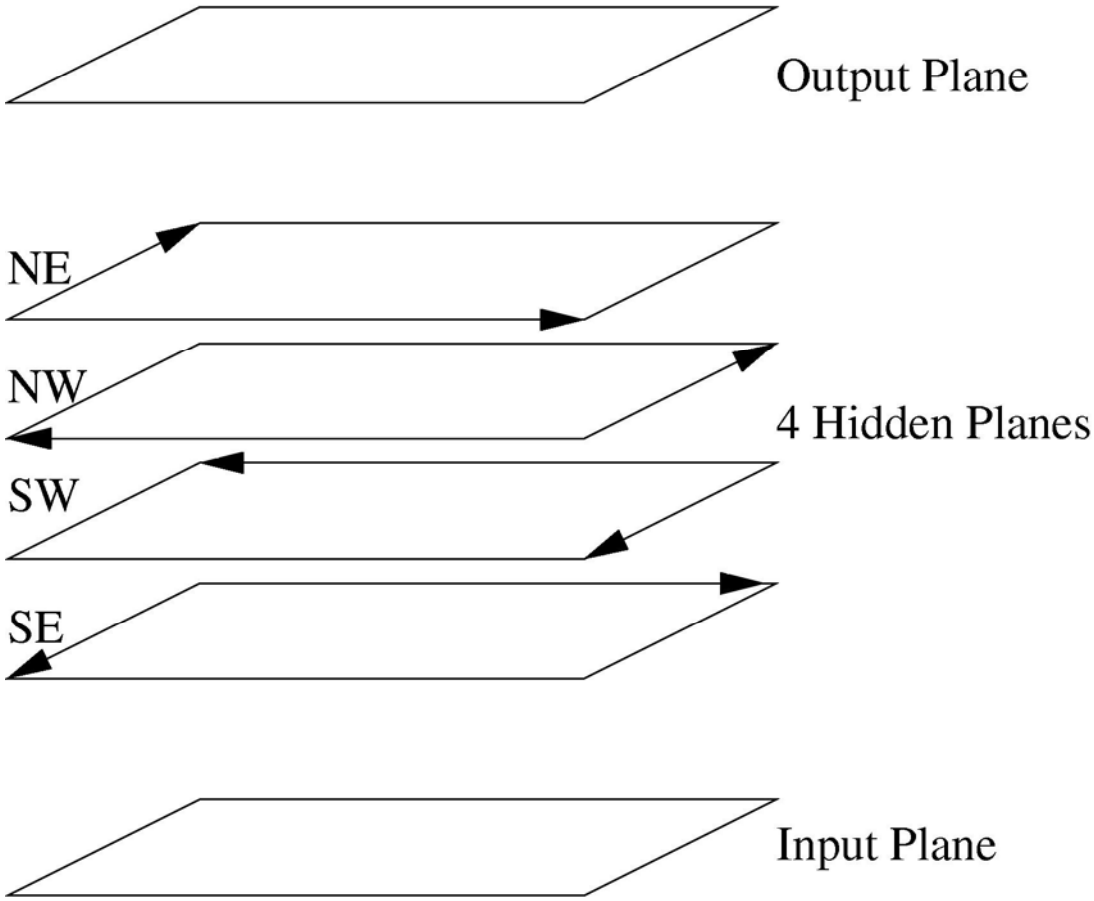


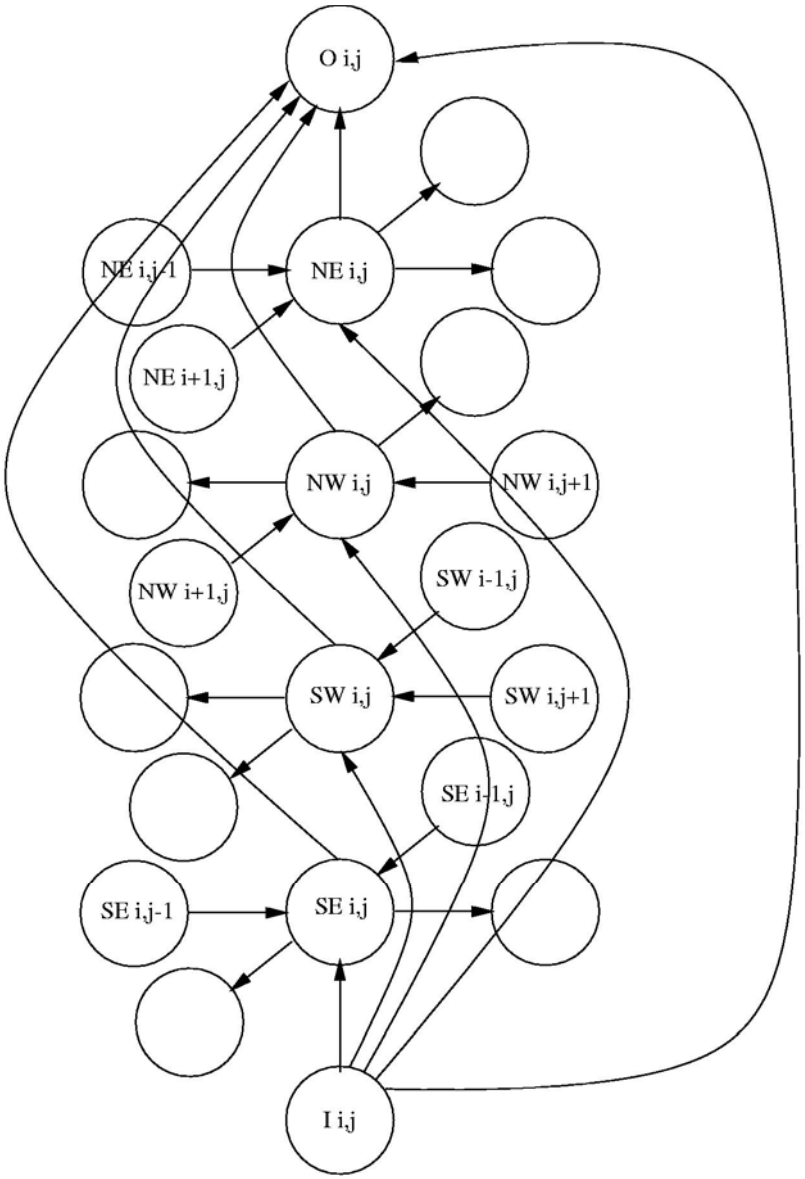

100
95
90
85
80
75
70
65
60
% correct prediction
0
10
20
30
40
50
60
70
80
90
100
Threshold (% exposed)

# FUNDAMENTAL LIMITATIONS

**100% CORRECT RECOGNITION IS PROBABLY IMPOSSIBLE FOR SEVERAL REASONS**

- **SOME PROTEINS DO NOT FOLD SPONTANEOUSLY OR MAY NEED CHAPERONES**
- **QUATERNARY STRUCTURE [BETA-STRAND PARTNERS MAY BE ON A DIFFERENT CHAIN]**
- **STRUCTURE MAY DEPEND ON OTHER VARIABLES [ENVIRONMENT, PH]**
- **DYNAMICAL ASPECTS**
- **FUZZINESS OF DEFINITIONS AND ERRORS IN DATABASES**


Output Plane
NE
NW
4 Hidden Planes
SW
SE
Input Plane

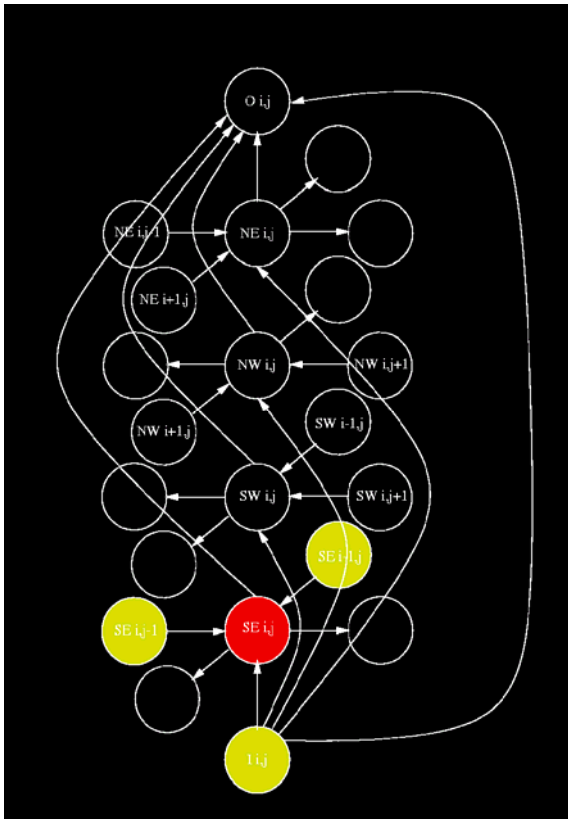

O i,j
NE i,j-1
NE i,j
NE i+1,j
NW i,j
NW i,j+1
NW i+1,j
SW i-1,j
SW i,j
SW i,j+1
SE i-1,j
SE i,j-1
SE i,j
I i,j

# BB-RNNs

$$
\begin{cases}
O_{ij} = \mathcal{N}_O(I_{ij}, H_{i,j}^{NW}, H_{i,j}^{NE}, H_{i,j}^{SW}, H_{i,j,}^{SE}) \\
H_{i,j}^{NE} = \mathcal{N}_{NE}(I_{i,j}, H_{i-1,j}^{NE}, H_{i,j-1}^{NE}) \\
H_{i,j}^{NW} = \mathcal{N}_{NW}(I_{i,j}, H_{i+1,j}^{NW}, H_{i,j-1}^{NW}) \\
H_{i,j}^{SW} = \mathcal{N}_{SW}(I_{i,j}, H_{i+1,j}^{SW}, H_{i,j+1}^{SW}) \\
H_{i,j}^{SE} = \mathcal{N}_{SE}(I_{i,j}, H_{i-1,j}^{SE}, H_{i,j+1}^{SE})
\end{cases}
$$

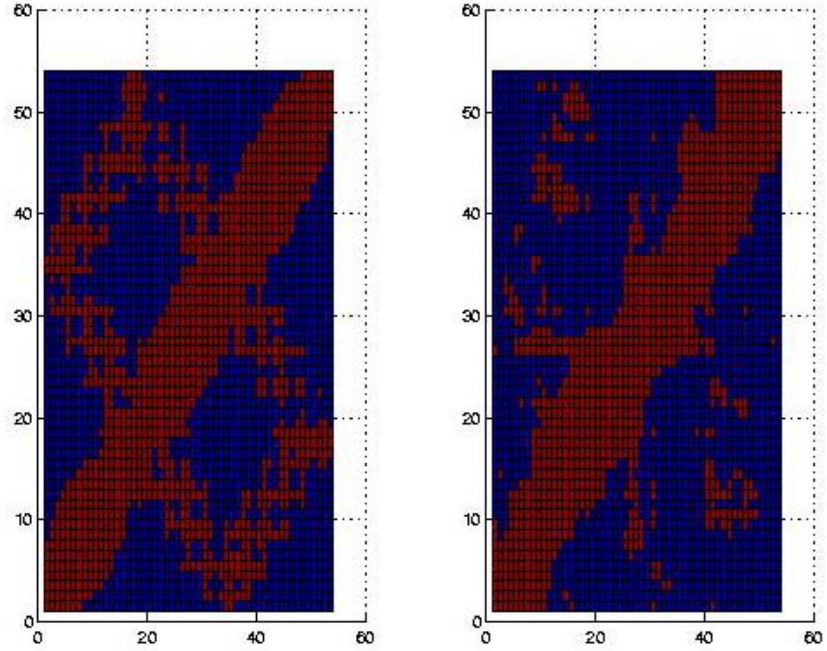
# 2D RNNs

$$
\begin{cases}
O_{ij} = \mathcal{N}_O(I_{ij}, H^{NW}_{i,j}, H^{NE}_{i,j}, H^{SW}_{i,j}, H^{SE}_{i,j,}) \\
H^{NE}_{i,j} = \mathcal{N}_{NE}(I_{i,j}, H^{NE}_{i-1,j}, H^{NE}_{i,j-1}) \\
H^{NW}_{i,j} = \mathcal{N}_{NW}(I_{i,j}, H^{NW}_{i+1,j}, H^{NW}_{i,j-1}) \\
H^{SW}_{i,j} = \mathcal{N}_{SW}(I_{i,j}, H^{SW}_{i+1,j}, H^{SW}_{i,j+1}) \\
H^{SE}_{i,j} = \mathcal{N}_{SE}(I_{i,j}, H^{SE}_{i-1,j}, H^{SE}_{i,j+1})
\end{cases}
$$



# 2D INPUTS

- **AA at positions i and j**
- **Profiles at positions i and j**
- **Correlated profiles at positions i and j**
- **+ Secondary Structure, Accessibility, etc.**


60
50
40
30
20
10
0
0
20
40
60
60
50
40
30
20
10
0
0
20
40
60

# PERFORMANCE (%)

| | 6Å | 8Å | 10Å | 12Å |
|---|---|---|---|---|
| **non-contacts** | **99.9** | **99.8** | **99.2** | **98.9** |
| **contacts** | **71.2** | **65.3** | **52.2** | **46.6** |
| **all** | **98.5** | **97.1** | **93.2** | **88.5** |

# A Perfectly Predicted Example

**Sequence with cysteine's position identified:**

MSNHTHHLKFKTLKRAWKASKYFIVGLSC[29]LYKFNLKSLVQTALSTLAMITLTSLVITAIIYISVGN
AKAKPTSKPTIQQTQQPQNHTSPFFTEHNYKSTHTSIQSTTLSQLLNIDTTRGITYGHSTNETQNR
KIKGQSTLPATRKPPINPSGSIPPENHQDHNNFQTLPYVPC[173]STC[176]EGNLAC[182]LSLC[18

6]HIETERAPSRAPTITLKKTPKPKTTKKPTKTTIHHRTSPETKLQPKNNTATPQQG
ILSSTEHHTNQSTTQI

Length: 257, Total number of cysteines: 5

Four bonded cysteines form two disulfide bonds :

173 -------186 ( red cysteine pair)

176 -------182 (blue cysteine pair)

**Prediction Results from DIpro (http://contact.ics.uci.edu/bridge.html)**

Predicted Bonded Cysteines:

173,176,182,186

Predicted disulfide bonds

| Bond_Index | Cys1_Position | Cys2_Position |
|---|---|---|
| 1 | 173 | 186 |
| 2 | 176 | 182 |

Prediction Accuracy for both bond state and bond pair are 100%.

# A Hard Example with Many Non-Bonded Cysteines

## Sequence with cysteine's position identified:

MTLGRRLAC[9]LFLAC[14]VLPALLLGGTALASEIVGGRRARPHAWPFMVSLQLRGGHFC
[55]GATLIAPNFVMSAAHC[71]VANVNVRAVRVVLGAHNLSRREPTRQVFAVQRIFENGY
DPVNLLNDIVILQLNGSATINANVQVAQLPAQGRRLGNGVQC[151]LAMGWGLLGRNRGI
ASVLQELNVTVVTSLC[181]RRSNVC[187]TLVRGRQAGVC[198]FGDSGSPLVC[208]NG
LIHGIASFVRGGC[223]ASGLYPDAFAPVAQFVNWIDSIIQRSEDNPC[254]PHPRDPDPAS
RTH

Length: 267, Total Cysteine Number: 11

Eight bonded cysteines form four disulfide bonds:

55 ----- 71 (Red), 151 ----- 208 (Blue), 181 ----- 187 (Green), 198 ----- 223 (Purple)

## Prediction Results from DIpro (http://contact.ics.uci.edu/bridge.html)

Predicted Bonded Cysteines:

9,14,55,71,181,187,223,254

Predicted Disulfide Bonds:

| Bond_Index | Cys1_Position | Cys2_Position | |
|---|---|---|---|
| 1 | 55 | 71 | (correct) |
| 2 | 9 | 14 | (wrong) |
| 3 | 223 | 254 | (wrong) |
| 4 | 181 | 187 | (correct) |

Bond State Recall: 5 / 8 = 0.625, Bond State Precision = 5 / 8 = 0.625

Pair Recall = 2 / 4 = 0.5 ; Pair Precision = 2 / 4 = 0.5

Bond number is predicted correctly.

# Prediction Accuracy on SP51 Dataset on All Cysteines

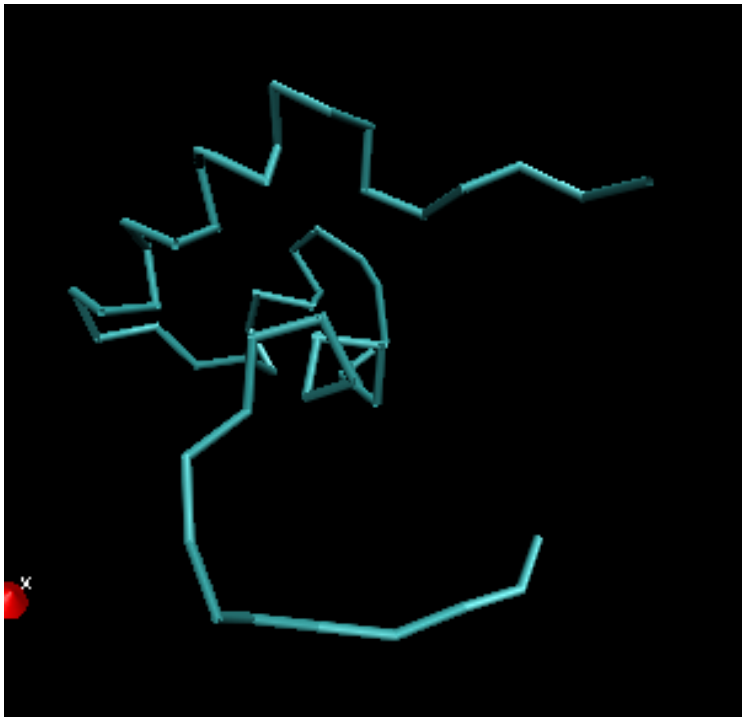
| Bond Num | Bond State Recall(%) | Bond State Precision(%) | Pair Recall(%) | Pair Precision(%) |
|---|---|---|---|---|
| 1 | 91 | 46 | 74 | 39 |
| 2 | 93 | 77 | 61 | 51 |
| 3 | 90 | 74 | 54 | 45 |
| 4 | 77 | 87 | 52 | 59 |
| 5 | 71 | 86 | 33 | 42 |
| 6 | 65 | 84 | 27 | 34 |
| 7 | 63 | 85 | 36 | 55 |
| 8 | 66 | 89 | 27 | 41 |
| 9 | 60 | 83 | 23 | 35 |
| 10 | 55 | 86 | 30 | 45 |
| 11 | 62 | 86 | 34 | 47 |
| 12 | 67 | 97 | 17 | 23 |
| 15 | 50 | 94 | 27 | 50 |
| 16 | 82 | 99 | 11 | 13 |
| 17 | 61 | 96 | 22 | 33 |
| 18 | 50 | 82 | 6 | 9 |
| 19 | 47 | 90 | 11 | 20 |

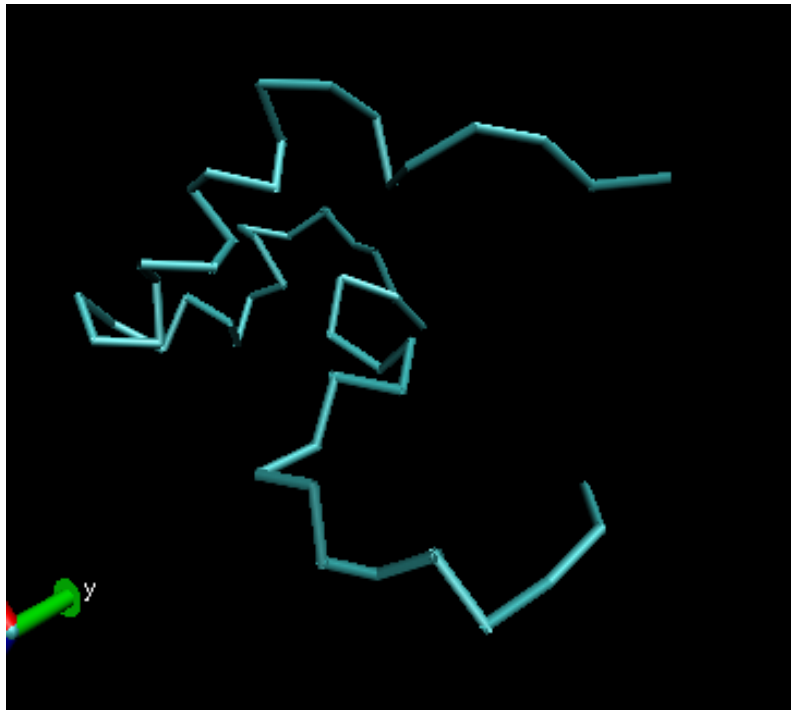
Overall bond state recall: 78%; overall bond state precision: 74%;
bond number prediction accuracy: 53%;
average difference between true bond number and predicted bond number: 1.1 .

# Protein Reconstruction

Using predicted secondary structure and predicted contact map

```
PDB ID   : 1HCR, chain A
Sequence: GRPRAINKHEQEQISRLLEKGHPRQQLAIIFGIGVSTLYRYFPASSIKKRMN
True SS : CCCCCCCCHHHHHHHHHHHHCCCCHHHHHHHCECCHHHHHHHCCCCCCCCCCC
Pred SS : CCCCCCCHHHHHHHHHHHHHCCCCHHHHEEHECHHHHHHHHCCCHHHHHHHCC
```





PDB ID: 1HCR

Chain A (52 residues)

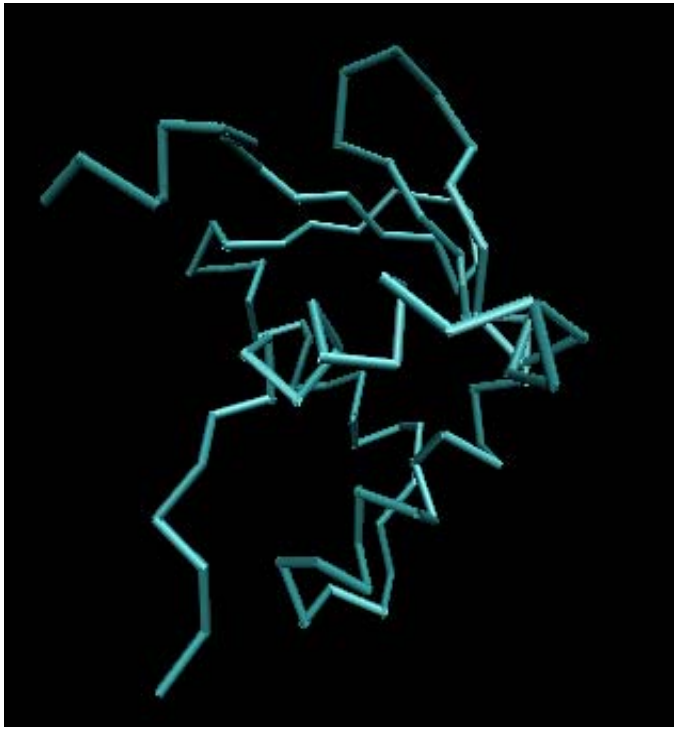
Model # 147

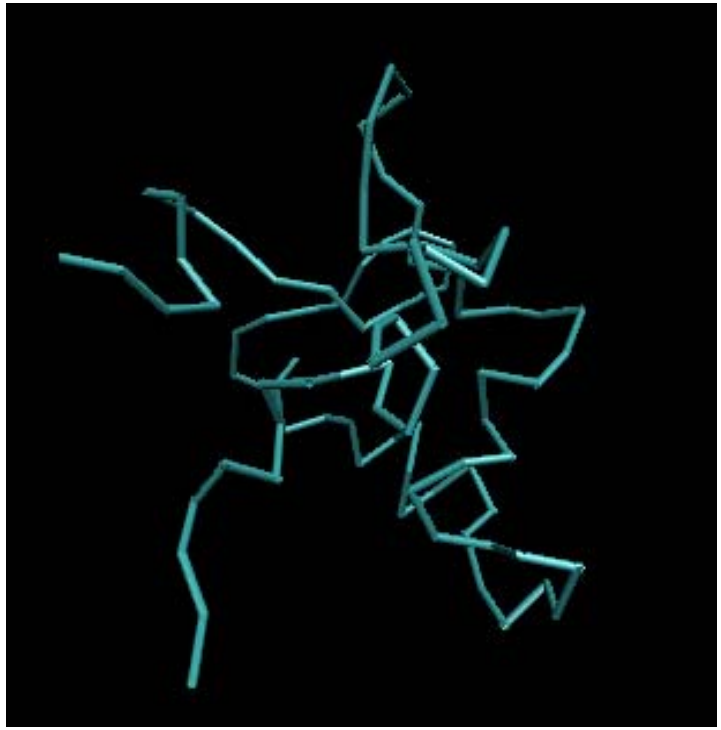
RMSD  3.47Å

# Protein Reconstruction

## Using predicted secondary structure and predicted contact map

```
PDB ID   : 1BC8, chain C
Sequence: MDSAITLWQFLLQLLQKPQNKHMICWTSNDGQFKLLQAEEVARLWGIRKNKPNMNYDKLSRALRYYYVKNIIKKVNGQKFVYKFVSYPEILNM
True SS : CCCCCCHHHHHHHHCCCHHHCCCCEECCCCCEEECCHHHHHHHHHHHHCCCCCCHHHHHHHHHHHHHHCCEEECCCCCCEEEECCCCHHHCC
Pred SS : CCCHHHHHHHHHHHHHCCCCCCEEEEECCCEEEEECCHHHHHHHHHHHCCCCCCCHHHHHHHHHHHHHCCCEEECCCCEEEEEEECCHHHHCC
```





PDB ID: 1BC8

Chain C (93 residues)

Model # 1714

RMSD  4.21Å

# CURRENT WORK

- **Feedback:**

**Ex: SS → Contacts → SS → Contacts**

- **Homology, homology, homology:**

**SSpro 4.0 performs at 88%**

# STRUCTURAL PROTEOMICS SUITE

**www.igb.uci.edu**

- SSpro: secondary structure
- SSpro8: secondary structure
- ACCpro: accessibility
- CONpro: contact number
- DI-pro: disulphide bridges
- BETA-pro: beta partners
- CMAP-pro: contact map
- CCMAP-pro: coarse contact map
- CON23D-pro: contact map to 3D
- 3D-pro: 3D structure

# SSpro 2, SSpro8, ACCpro, CONpro

(Server statistics, Update history, Quick help and References)

- **SSpro 2**: Protein secondary structure prediction based on Bidirectional Recurrent Neural Networks (BRNNs). Improved performance with the new PSI-BLAST profiles.
- **SSpro8**: Experimental 8-class secondary structure prediction based on BRNNs.
- **CONpro**: Prediction of number of residue contacts in a radius of 6, 8, 10 and 12 Å based on BRNNs.
- **ACCpro**: Prediction of residue solvent accessibility at variable thresholds (0%-75%) based on BRNNs.

Your email address (where the prediction will be sent):

Name of your query (optional):

Predictions:

- [x] SSpro 2
- [ ] SSpro8
- [ ] CONpro
- [ ] ACCpro - Threshold: 25%

Paste your protein sequence here (plain sequence, no headers, spaces and newlines will be ignored):

Predict

Attention: due to BLAST and PSI-BLAST searches on the NR database it may take several minutes to serve a query.

---

*Gianluca Pollastri, gpollast@ics.uci.edu*

```
SISQQTVWNQMATVRTPLNFDSSKQSFCQFSVDLLGGGISVDKTGDWITLVQNSPISNLL
CCCECCCCCCEEEECCCCCCCCCCCCEEEEEEECCCCEEEECCCCCCEEEEECCHHHHHH
CCCEEEEECEEEEECCCCCCCTCCCCEEEEEEEETCSEEEECTTTTEEEEEECCHHHHHH
-----------+-------------+++++++++-+---------+++----++++++
---------+-+++----------+++++++++-+-++++---+++++++++++++++
----------+--++--------+++++++++---+++-----+++++++++++++
---------+-++++++--------+++++++++---+++-----+++++++++++++
eeeeee---e--e-e-eee-ee-eee---------e-e--eeeeee-------------


RVAAWKKGCLMVKVVMSGNAAVKRSDWASLVQVFLTNSNSTEHFDACRWTKSEPHSWELI
HHHHHHCCCEEEEEEEEEECCEEECCCCCEEEEEEEECCCCCCCCCEEEEEECCCCCCCC
HHHHHHTTCEEEEEEEEEEEEEECCCCCEEEEEEEECCCTTCCCEEEEEEECCTCCEEE
+++++--++++++++++-+----------++++++---------+-+++----------
+++----++++++++++++----------+++++++++------+++++++++-+-+--
++++---++++++++++++--+------+++++++++------+++++++++---+-+
++++-+-++++++++++++----------+++++++++------+++++++++---+-+
-----ee---e-------e-e-ee-e-e-e-----e--eeee--e-------e-e-ee-e

..

Solvent accessibility threshold: 25%
PSI-BLAST hits : 24

..

Query served in 151 seconds
```

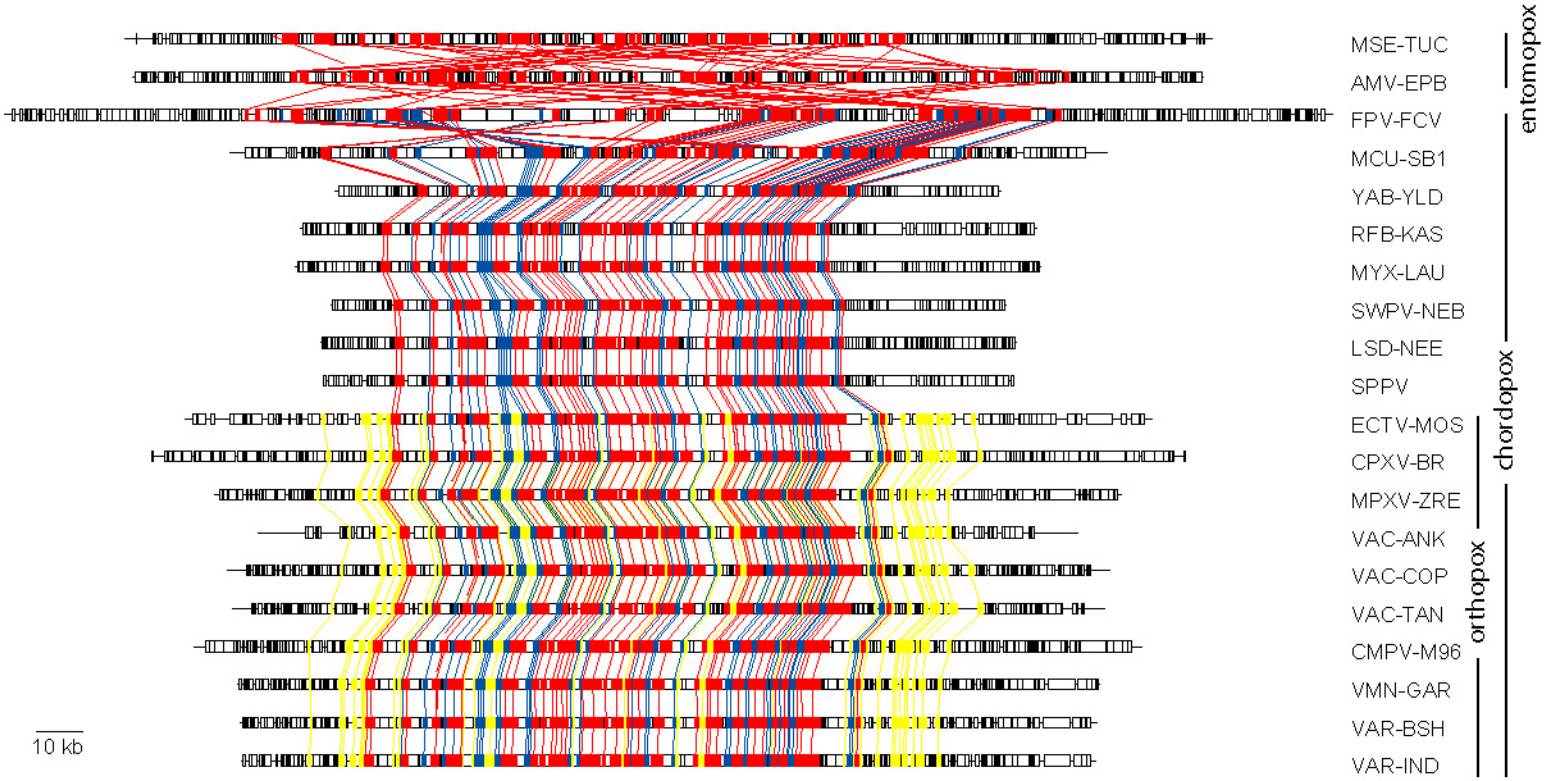
# Advantage of Machine Learning

- **Pitfalls of traditional ab-initio approaches**
- **Machine learning systems take time to train (weeks).**
- **Once trained however they can predict structures almost faster than proteins can fold.**
- **Predict or search protein structures on a genomic or bioengineering scale .**

# Structural Databases

- **PPDB = Poxvirus Proteomic Database**
- **ICBS = Inter Chain Beta Sheet Database**

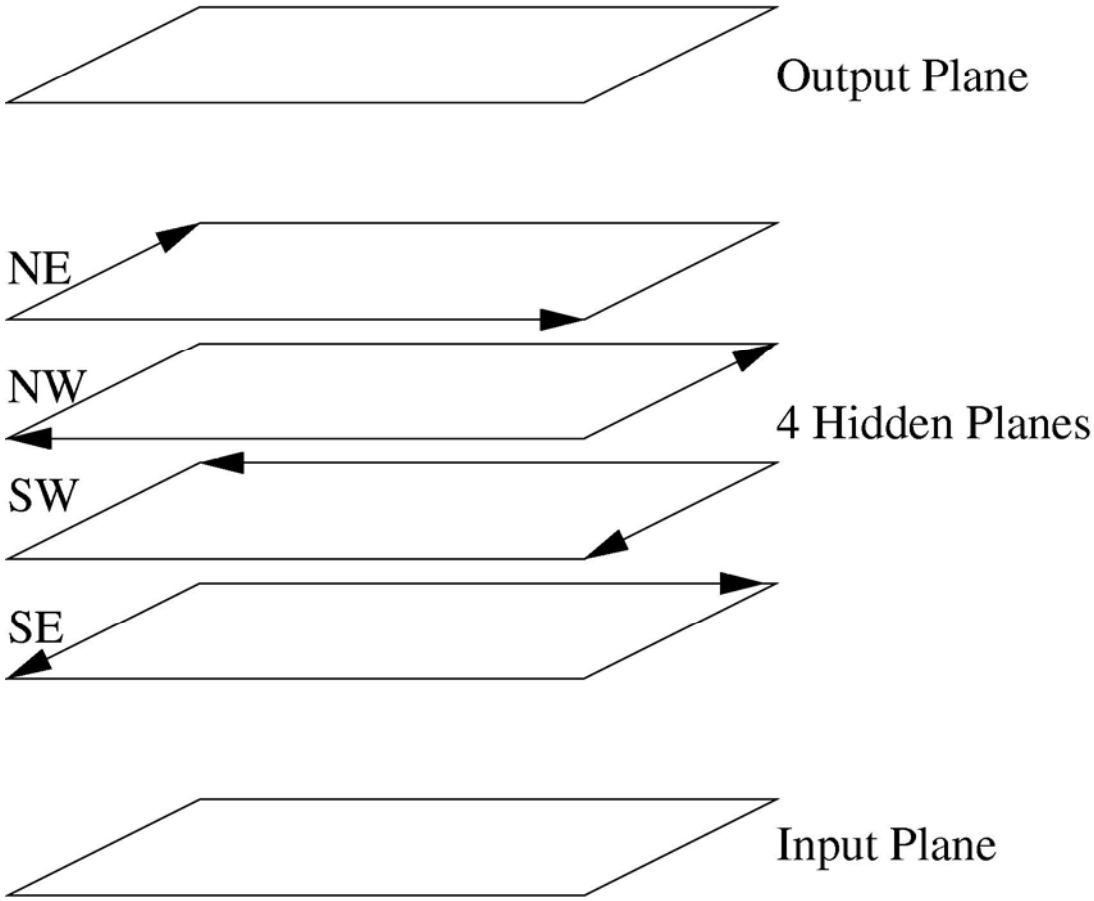
mclysaght_fig2

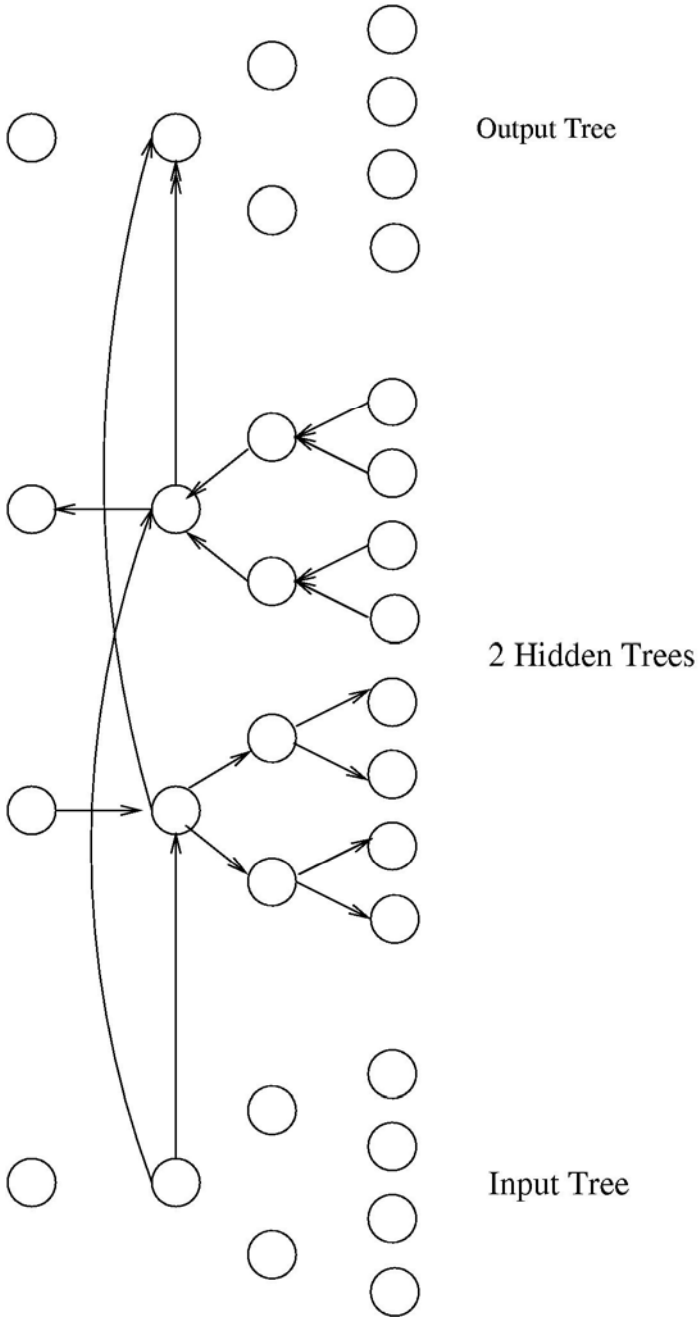


# DAG-RNNs APPROACH

- **Two steps:**
  - 1. Build relevant DAG to connect inputs, outputs, and hidden variables
  - 2. Use a deterministic (neural network) parameterization together with appropriate stationarity assumptions/weight sharing—overall models remains probabilistic

- **Process structured data of variable size, topology, and dimensions efficiently**
- **Sequences, trees, d-lattices, graphs, etc**
- **Convergence theorems**
- **Other applications**


Output Plane
NE
NW
4 Hidden Planes
SW
SE
Input Plane


Output Tree
2 Hidden Trees
Input Tree

# Convergence Theorems

- **Posterior Marginals:**

**σBN→dBN in distribution**

**σBN→dBN in probability (uniformly)**

- **Belief Propagation:**

**σBN→dBN in distribution**

**σBN→dBN in probability (uniformly)**

# ACKNOWLEDGMENTS


- **UCI:**
  - Gianluca Pollastri, Michal Rosen-Zvi
  - Arlo Randall, Pierre-Francois Baisnee, S. Josh Swamidass, Jianlin Cheng, Yimeng Dou, Yann Pecout, Mike Sweredoski, Alessandro Vullo, Lin Wu,
  - James Nowick, Luis Villareal
- **DTU:** Soren Brunak
- **Columbia:** Burkhard Rost
- **U of Florence:** Paolo Frasconi
- **U of Bologna:** Rita Casadio, Piero Fariselli


www.igb.uci.edu/tools.htm
www.ics.uci.edu/~pfbaldi