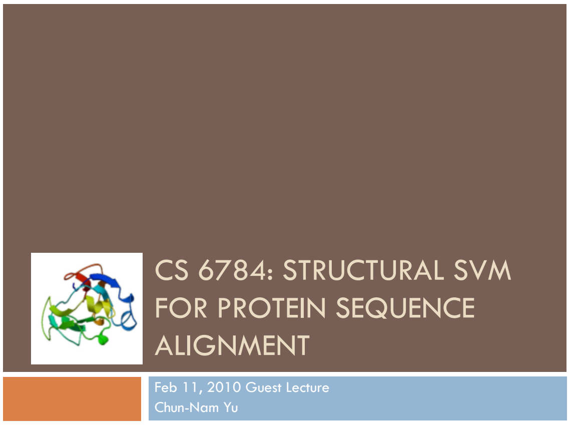
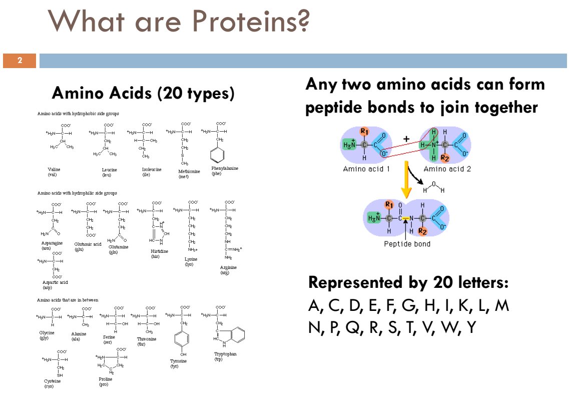


CS 6784: STRUCTURAL SVM FOR PROTEIN SEQUENCE ALIGNMENT

Feb 11, 2010 Guest Lecture
Chun-Nam Yu



What are Proteins?

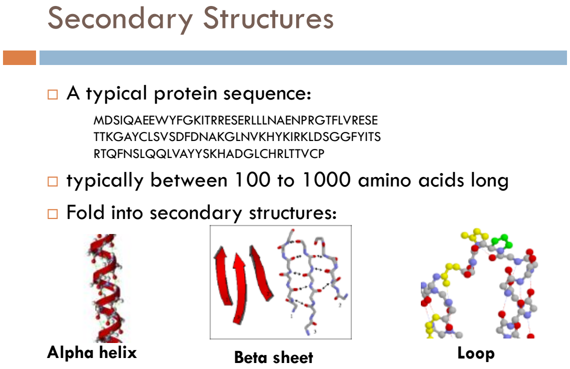
2

Amino Acids (20 types)

Any two amino acids can form peptide bonds to join together

Represented by 20 letters:

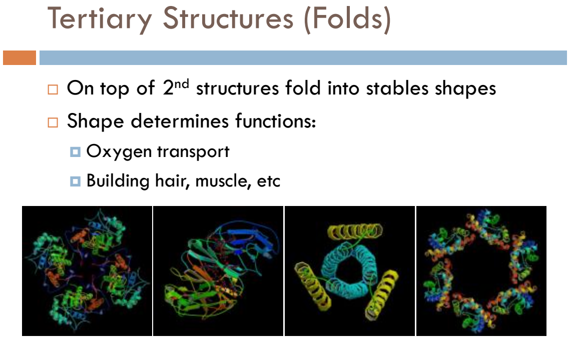
A, C, D, E, F, G, H, I, K, L, M
N, P, Q, R, S, T, V, W, Y



Secondary Structures

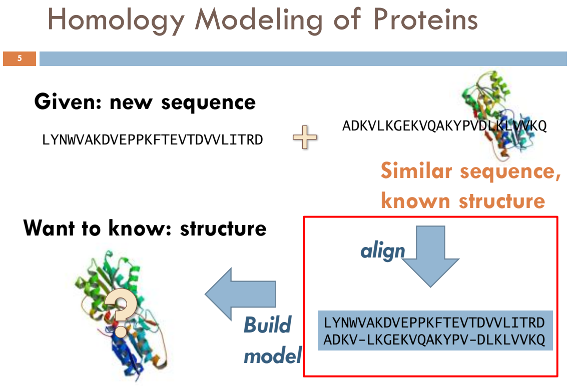
- A typical protein sequence:

  MDSIQAEEWYFGKITRRESERLLLNAENPRGTFLVRESE
  TTKGAYCLSVSDFDNAKGLNVKHYKIRKLDSGGFYITS
  RTQFNSLQQLVAYYSKHADGLCHRLTTVCP
- typically between 100 to 1000 amino acids long
- Fold into secondary structures:



Tertiary Structures (Folds)

- On top of 2nd structures fold into stables shapes
- Shape determines functions:
  - Oxygen transport
  - Building hair, muscle, etc



Homology Modeling of Proteins

5

Given: new sequence

LYNWVAKDVEPPKFTEVTDVVLITRD

ADKVLKGEKVQAKYPVDLKLVVKQ

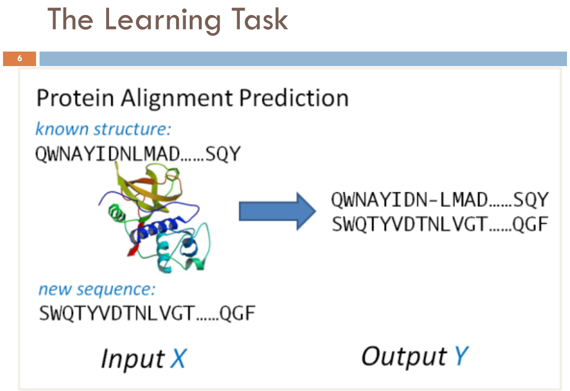
Similar sequence, known structure

Want to know: structure

align

LYNWVAKDVEPPKFTEVTDVVLITRD
ADKV-LKGEKVQAKYPV-DLKLVVKQ

Build model



The Learning Task

6

Protein Alignment Prediction

known structure:

QWNAYIDNLMAD……SQY

QWNAYIDN-LMAD……SQY
SWQTYVDTNLVGT……QGF

new sequence:

SWQTYVDTNLVGT……QGF

Input X

Output Y

Sequence Alignments

- □ Given sequence pairs:

AECD EACC

- □ Alignment 1:

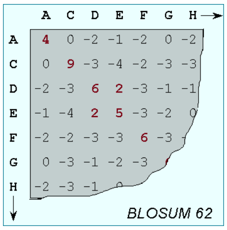
AECD
EACC

Score
= -1 – 1 +9 + 6
= 13

- □ Alignment 2:

-AECD
EACC-

Score
= -3 + 4 – 4 + 9 - 3
= 3



	A	C	D	E	F	G	H
A	4	0	-2	-1	-2	0	-2
C	0	9	-3	-4	-2	-3	-3
D	-2	-3	6	2	-3	-1	-1
E	-1	-4	2	5	-3	-2	0
F	-2	-2	-3	-3	6	-3	-
G	0	-3	-1	-2	-3		
H	-2	-3	-1	0			

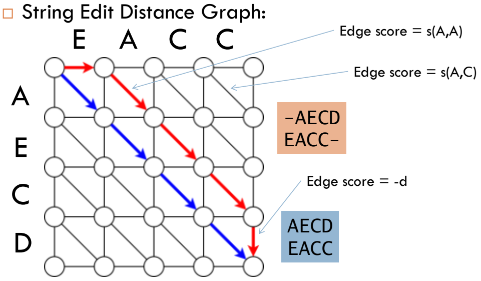
BLOSUM 62

Substitution Cost Matrix

Gap penalty *d* = -3

Smith-Waterman Algorithm

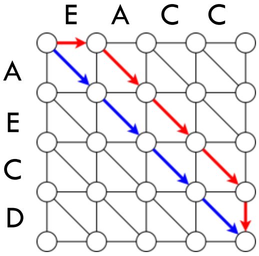
- □ Weighted string edit distance: Match, insert, delete
- □ String Edit Distance Graph:



Smith-Waterman Algorithm

- □ Weighted string edit distance: Match, insert, delete
- □ Find highest scoring path through the graph
- □ Dynamic Programming Recurrence:

$$F(i,j) = \max\begin{cases} F(i-1,j-1) + s(a_i,b_j) \\ F(i-1,j) - d \\ F(i,j-1) - d \end{cases}$$

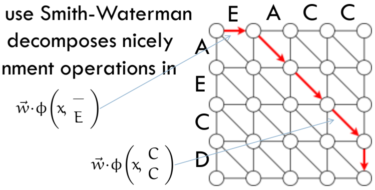


More Complex Substitution Scores

- □ Consider More General Linear Scoring Rule:

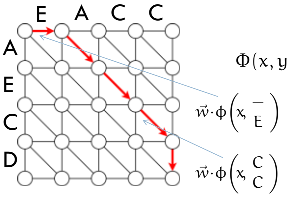
$$\Phi(x,y) = \phi\left(x, \begin{matrix} - \\ E \end{matrix}\right) + \phi\left(x, \begin{matrix} A \\ A \end{matrix}\right) + \phi\left(x, \begin{matrix} E \\ C \end{matrix}\right) + \phi\left(x, \begin{matrix} C \\ C \end{matrix}\right) + \phi\left(x, \begin{matrix} D \\ - \end{matrix}\right)$$

- □ Score of alignment = $\vec{w} \cdot \Phi(x,y)$
- □ Can still use Smith-Waterman as score decomposes nicely into alignment operations in y


E A C C
A
E
C
D
$\vec{w} \cdot \phi\left(x, \begin{matrix} - \\ E \end{matrix}\right)$
$\vec{w} \cdot \phi\left(x, \begin{matrix} C \\ C \end{matrix}\right)$

More Complex Substitution Scores

- □ In the simplest case the match score $s(a_i, b_j)$ is just a simple lookup over the BLOSUM matrix
- □ Consider the feature map:
- □ Equivalent to BLOSUM


E A C C
A
E
C
D
$\vec{w} \cdot \phi\left(x, \begin{matrix} - \\ E \end{matrix}\right)$
$\vec{w} \cdot \phi\left(x, \begin{matrix} C \\ C \end{matrix}\right)$

$$\Phi(x,y) = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 1 \\ 0 \\ \vdots \\ 1 \\ 0 \\ \vdots \\ 2 \end{pmatrix}$$

Align 'A' with 'A'

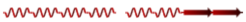
Align 'C' with 'C'

Align 'E' with 'C'

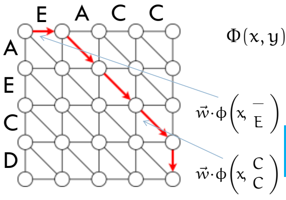
Gap

More Complex Substitution Scores

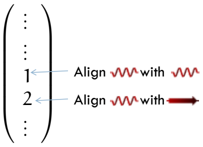
- □ Suppose we also have information on 2nd structures:
- □ x = (A E C D, E A C C)



- □ Extra feature counts:


E A C C
A
E
C
D
$\vec{w} \cdot \phi\left(x, \begin{matrix} - \\ E \end{matrix}\right)$
$\vec{w} \cdot \phi\left(x, \begin{matrix} C \\ C \end{matrix}\right)$

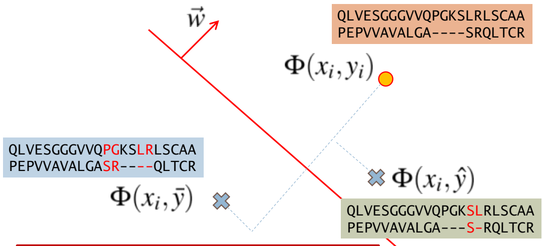
$\Phi(x,y) =$



Our substitution score includes 2nd structure information too!

Structural Support Vector Machines

13



for all alignments $\hat{y} \in \mathcal{Y}$

$$\vec{w} \cdot \Phi(x_i, y_i) - \vec{w} \cdot \Phi(x_i, \hat{y}) \geq \Delta(y_i, \hat{y}) - \xi_i$$

Structural Support Vector Machines

14

- Structural SVM [Tsochantaridis et al. '04]

$$\min_{\vec{w},\vec{\xi}} \frac{1}{2}\|\vec{w}\|^2 + C\sum_{i=1}^{n} \xi_i$$

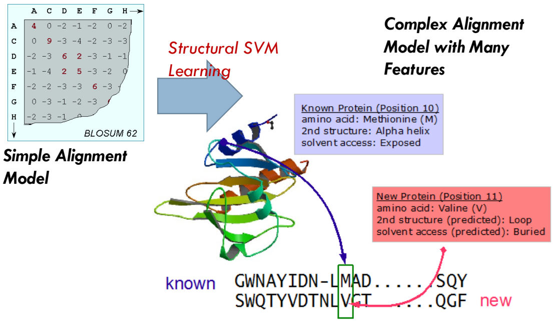
s.t. for $1 \cdot i \cdot n$, for all alignments $\hat{y} \in \mathcal{Y}$,

$$\vec{w} \cdot \Phi(x_i, y_i) - \vec{w} \cdot \Phi(x_i, \hat{y}) \geq \Delta(y_i, \hat{y}) - \xi_i$$

- Convex optimization problem
- $O(\binom{n+m}{m})$ = exponentially many constraints
- Can solve using cutting-plane algorithm

Learning with Many Features

15



Simple Alignment Model

Structural SVM Learning

Complex Alignment Model with Many Features

Known Protein (Position 10)
amino acid: Methionine (M)
2nd structure: Alpha helix
solvent access: Exposed

New Protein (Position 11)
amino acid: Valine (V)
2nd structure (predicted): Loop
solvent access (predicted): Buried

```
known   GWNAYIDN-LMAD...../SQY
        SWQTYVDTNLVCT    ....QGF   new
```

Feature Vectors (2 examples)

- 3 basic structural features:
  - Amino acid (i.e., A, N, P, etc)
  - Secondary structures (i.e., ∿∿, ➡, ⎍)
  - Exposure to water (i.e., 1, 2, 3, 4, 5)
- Anova2:
  - Pairwise feature interaction, e.g., s(A and ∿∿, E and ➡)
- Window:
  - Consider neighborhood of aligned site, e.g., s(AEC, CED), s(∿∿ ∿∿ ∿∿ , ∿∿ ∿∿ ➡)

Loss Functions

Q - loss

- Correct Alignment y:

```
-AECD
EACC-
```

- Incorrect Alignment y':

```
A-ECD
EACC-
```

- Q-loss = 1/3

Q4 – loss (shift < 4)

- Correct Alignment y:

```
-AECD
EACC-
```

- Incorrect Alignment y':

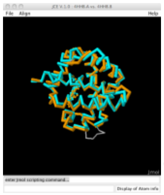
```
A-ECD
EACC-
```

- Q4-loss = 0

Experiments

18

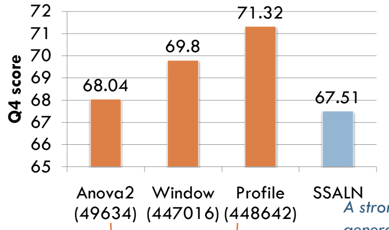
- Training Set: ~5000 alignments [Qiu & Elber '06]
- Test Set: ~30000 alignments from deposits to Protein Data Bank between June 05 to June 06
- All structural alignments produced by the program CE by superposition of 3D coordinates



(from pdb.org)

Results on Alignment Accuracy

- From [Yu, Joachims, Elber & Pillardy. RECOMB'07]



Structural SVM

A strong generative model with hand-tuned features

Summary

- An application of Structural SVM to a problem in computational biology
- Discriminative training allows us to incorporate complex features into the alignment models to improve alignment accuracy
- Showed how to design the feature vectors, loss functions, and inference procedures