# A Probabilistic Model for Deriving Structure of Proteins from Their Sequence Information

Alexey Nikolaev

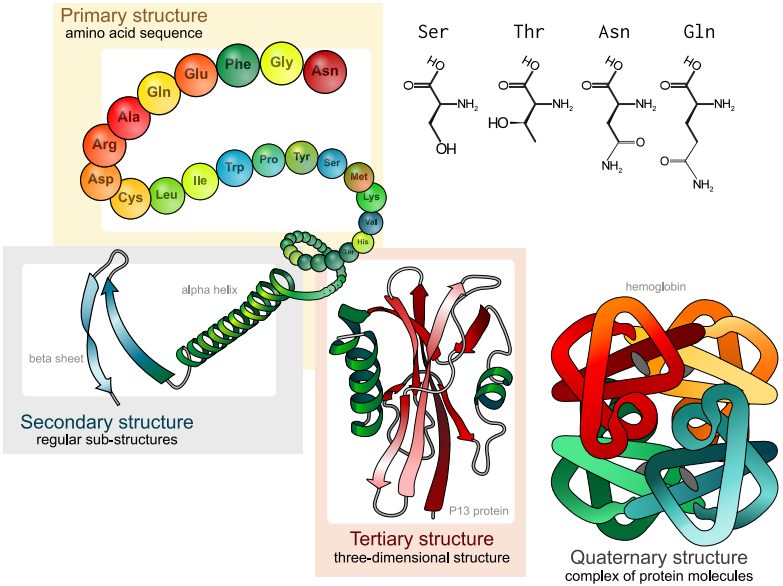
(The Graduate Center, CUNY)

Saad Mneimneh

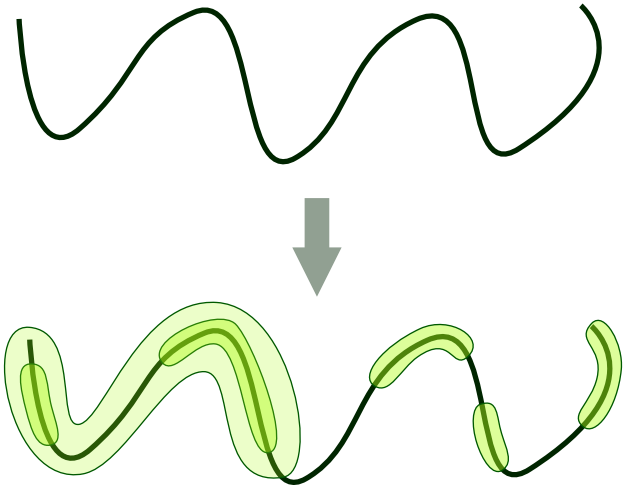
(Hunter College and the Graduate Center, CUNY)

March 22, 2013

# Protein Structure



# Can we find probably structurally important segments in a sequence?



# Clusters in a sequence of nodes

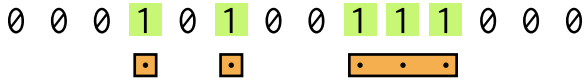


Assuming that some of the nodes in the sequence promote cluster formation, consider the following system:

There is a sequence of *1*s and *0*s.

0 0 0 1 0 1 0 0 1 1 1 0 0 0

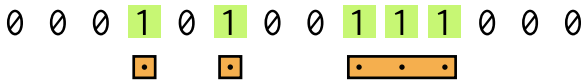
*1*s form clusters, while *0*s do nothing.



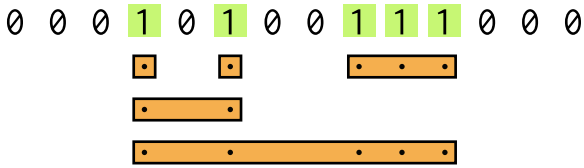
# Clusters in a sequence of nodes



Connect only immediate neighbors.



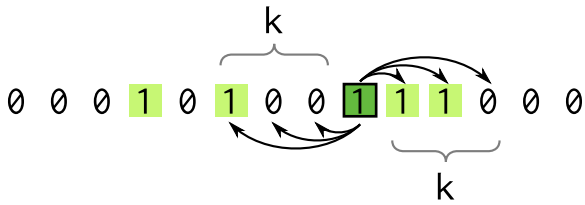

Is it good enough? Not really. We would like to capture clusters separated by *0*s.



# Clusters in a sequence of nodes

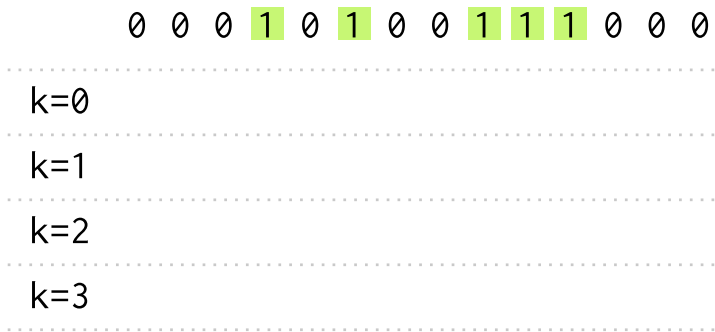


*Generalize:* Each node is connected to $k$ many nodes to the right, and $k$ many to the left. $k \geq 0$.

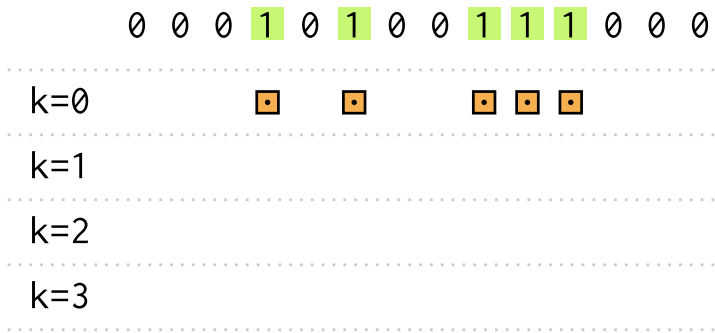



The resulting clusers *may have gaps* of at most $k - 1$ consecutive *0*s.

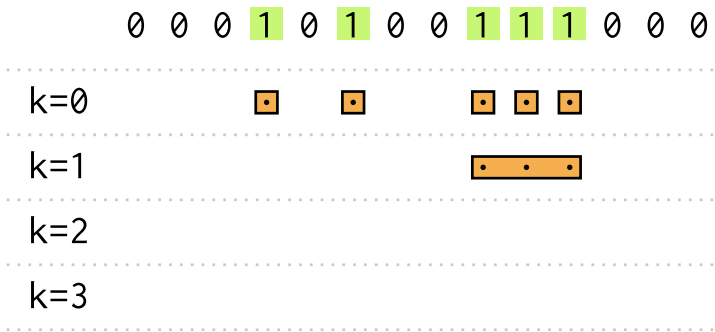
# Clusters for different $k$



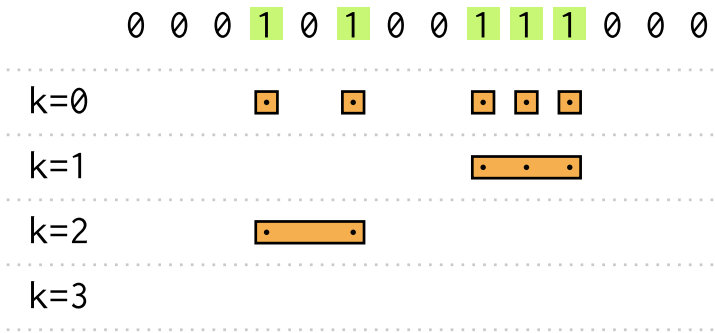
# Clusters for different $k$



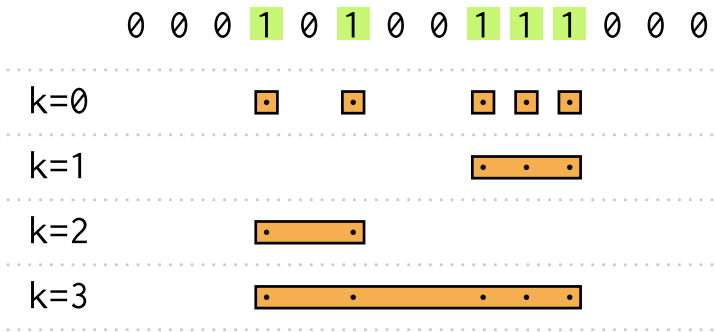
# Clusters for different $k$



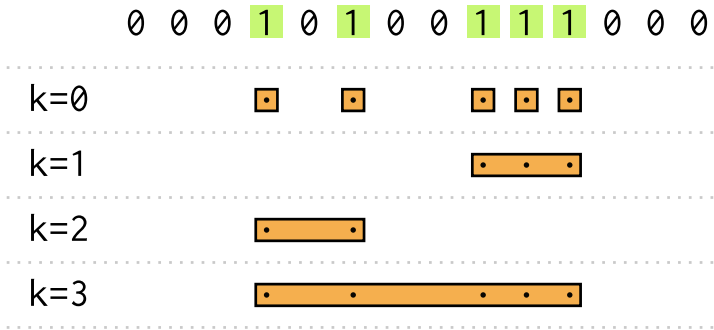
# Clusters for different $k$



# Clusters for different $k$



# Clusters for different $k$




Too many clusters! Which are really important?

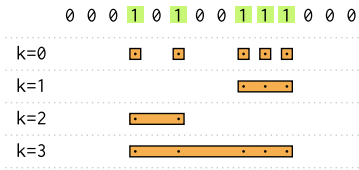
# Probabilistic model



If it is observed that *1*s and *0*s are found in sequences with certain probabilities:

- $p$ is the probability of *1*s, and
- $q = 1 - p$ is the probability of *0*s,

we can compute, how probable each of the clusters is.

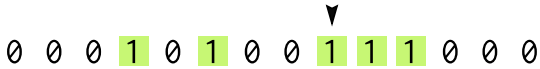


# Probabilistic model



**Def.** *Size of a cluster is the number of 1s in it.*

**Def.** *Given a 1, let $w_{k,s}$ be the probability to find that 1 in a cluster of size $s$ at level $k$.*




$$w_{k,s} = (\beta_{k,s} - \beta_{k-1,s}) \cdot q^{2k},$$

$$\text{where } \beta_{k,s} = s(p\alpha_k)^{s-1}, \text{ and } \alpha_k = \frac{1-q^k}{1-q}.$$

# Choosing the best cluster



Ok, if we found a cluster, how rare is it?

**Def.** Weight *of a cluster with size* $s$ *at level* $k$ *is*

$$W(k,s) = \frac{1}{\zeta_k} \min\left( \sum_{t=1}^{s} w_{k,t}, \sum_{t=s}^{\infty} w_{k,t} \right)$$

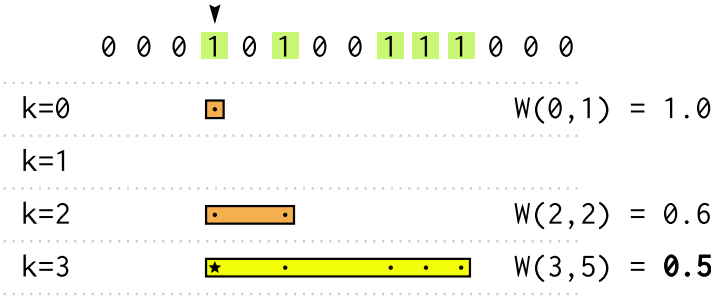
The normalizing constant $\zeta_k = \sum_{s=1}^{\infty} w_{k,s} = 1 - q^2$.

If a cluster has very small weight, it is not very likely to occure at random. Thus we can expect that it is important.

If a given *1* belongs to several clusters, we choose the one with the *least* weight.

# Choosing the best cluster

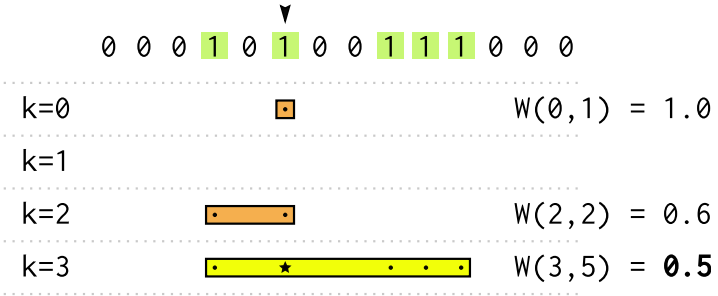





If a given *1* belongs to several clusters, we choose the one with the *least* weight.

# Choosing the best cluster

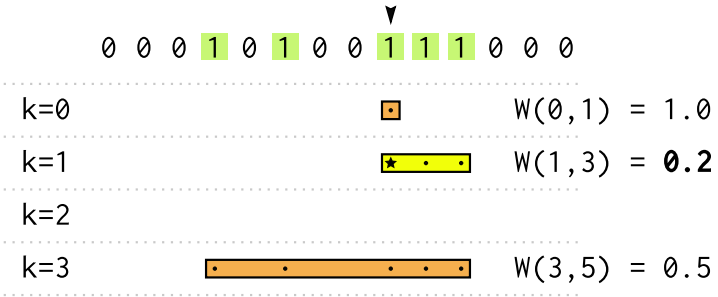





If a given *1* belongs to several clusters, we choose the one with the *least* weight.

# Choosing the best cluster

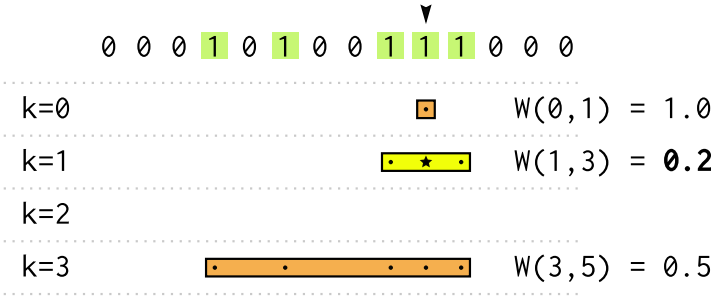





If a given *1* belongs to several clusters, we choose the one with the *least* weight.

# Choosing the best cluster

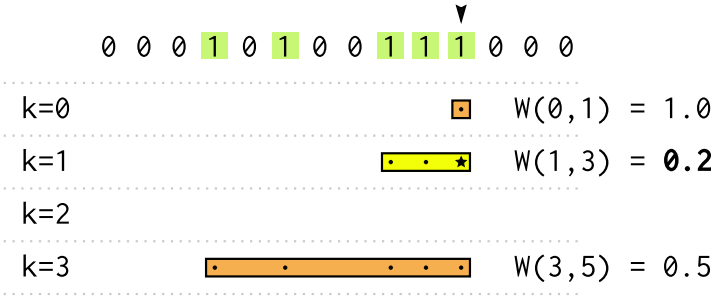





If a given *1* belongs to several clusters, we choose the one with the *least* weight.

# Choosing the best cluster

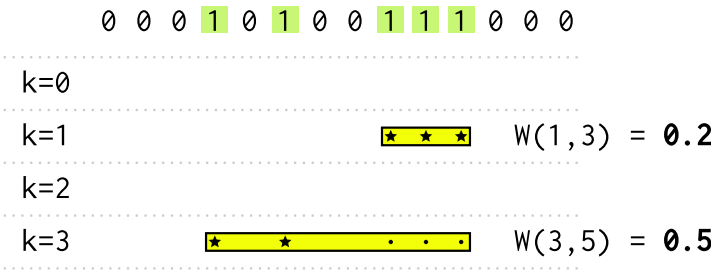





If a given *1* belongs to several clusters, we choose the one with the *least* weight.

# Chosen best clusters can be nested






If a given *1* belongs to several clusters, we choose the one with the *least* weight.

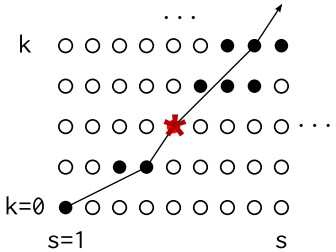
# Choosing the best cluster

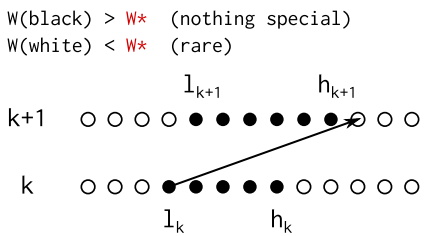


Let $P(k)$ be the probability that, for a given 1, the best cluster is at the level $k$.

**Theorem.** $P(k) = 0$ *for all* $k$.

That is, for any currently the best cluster, you can always find a better one, if the sequence is long enough.



# Choosing the best cluster



Given that up to level $k$ the best cluster has weight $W^*$, the probablity to find a new best cluster on the next level $k+1$ is

$$P_{upd!} = \sum_{i=\max(1,\, h_{k+1}+1-l_k)}^{\infty} t^i_{k,s} \geq \sum_{i=h_{k+1}}^{\infty} w_{k+1,i-1}\, p^2 = \sum_{i=h_{k+1}-1}^{\infty} w_{k+1,i}\, p^2$$

Notice that

$$W(k+1, h_{k+1}) \geq \frac{1}{1-q^2} \cdot \sum_{i=h_{k+1}}^{\infty} w_{k+1,i} \geq W^*$$

$$P_{upd!} > W^* (1-q^2)\, p^2 > 0.$$

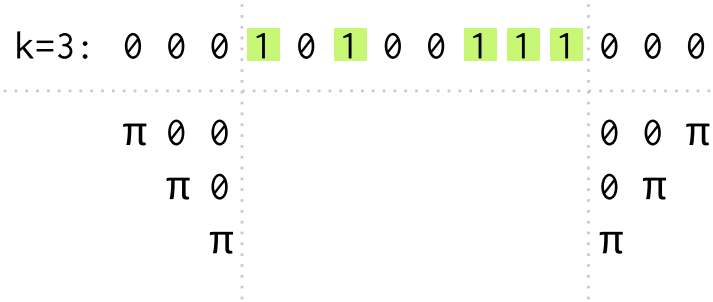
# Need for breakers



Some nodes that were previously zeroes now become *breakers*. Once reached, they stop cluster growth completely. Call them $\pi$ in our single-character notation.

Let $\pi$ also denote the probability of breakers.

$$p + q + \pi = 1$$



# Probability $w_{k,s}$ for the breakers case



With the introduction of breakers, we actually can get three types of clusters:

Open on both sides:

$$w_{k,s}^{(0\pi)} = (\beta_{k,s} - \beta_{k-1,s}) \cdot q^{2k}$$

With a breaker on one side:

$$w_{k,s}^{(1\pi)} = (\beta_{k,s} - \beta_{k-1,s}) \cdot 2q^{k} \alpha_k \pi$$

With breakers on both sides:

$$w_{k,s}^{(2\pi)} = (\beta_{k,s} - \beta_{k-1,s}) \cdot (\alpha_k \pi)^2$$

# Weight $W(k,s)$ for the breakers case



With the introduction of breakers, we actually can get three types of clusters:

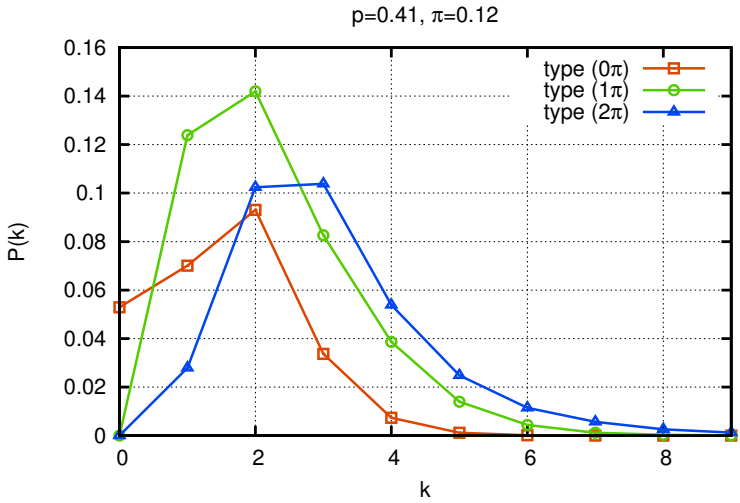
$$W^{(X\pi)}(k,s) = \frac{\min\left(\sum_{t=1}^{s} w_{k,t}^{(X\pi)}, \sum_{t=s}^{\infty} w_{k,t}^{(X\pi)}\right)}{\sum_{t=1}^{\infty}\left(w_{k,t}^{(0\pi)} + w_{k,t}^{(1\pi)} + w_{k,t}^{(2\pi)}\right)}$$

where $X \in \{0,\ 1,\ 2\}$.

$$\sum_{t=1}^{\infty} w_{k,t}^{(X\pi)} = \sum_{t=1}^{\infty}(\beta_{k,t} - \beta_{k-1,t}) \cdot C_k^{(X\pi)} = (B_k - B_{k-1}) \cdot C^{(X\pi)},$$

where $B_k = \dfrac{1}{(p\alpha_k - 1)^2}$, $C_k^{(0\pi)} = q^{2k}$, $C^{(1\pi)} = 2q^k\alpha_k\pi$, and $C_k^{(2\pi)} = (\alpha_k\pi)^2$. Also, $\alpha_k = (1-q^k)/(1-q)$ (the same as before).

# $P(k)$. The probability to choose a cluster at level $k$.



# Experiments with pretein databases



Can we make our method find secondary structures (helices and strands)?

How amino acids map to $\{1,\ 0,\ \pi\}$? Use genetic algorithm.

We simply say that if a residue is covered by any of our clusters, we predict that it belongs to a helix or a strand. Then, check, how good the prediction is.

$$\text{Fitness} = \frac{\text{number of correctly predicted residues}}{\text{total number of residues}}$$

# Experiments with pretein databases



We get with fitness 67%:

$$\{V,\ I,\ L,\ F,\ M,\ Y,\ W,\ A\} \to 1$$
$$\{P,\ G\} \to \pi$$
$$\text{others} \to 0$$

Hydrophobic amino acids are responsible for cluster formation.

Can we really predict secondary structures?

# Secondary structure prediction?



There are “Helix”, “Strand”, and “Coil” regions.

1) Drop clusters that have size $s = 1$.

2) We predict that residues in clusters formed at levels $k = 1$ and $k = 2$ are *Strands*.

3) We predict that the remaining residues in other clusters are *Helices*.

4) The rest residues are *Coils*.

$$\text{Q3} = \frac{\text{number of correctly predicted residues}}{\text{total number of residues}}$$

# Secondary structure prediction?



Genetic algorithm on randomly selected records from DSSP produced the following map:

$$\{V,\ I,\ L,\ F,\ M,\ Y\} \to 1$$
$$\{P,\ G\} \to \pi$$
$$\text{others} \to 0$$

With this map, on a standard protein dataset CB-513, we get

$$Q3 = 55\%.$$

This is not 70-80%, but still it is better than, e.g. Chou-Fasman method that has $Q3 = 46 - 48\%$.

# Future work



1. To go beyond secondary structures:
   - How to make breakers weaker?
   - Probabilistic assignment of the map residue $\rightarrow \{1, 0, \pi\}$.
   - Get rid of breakers, and insert strings of zeroes instead, e.g. $P \mapsto 00000$, and $G \mapsto 00$.
2. How far can we get in predicting sec. structures?
   - Map pairs or triples of residues to $\{1, 0, \pi\}$.
   - Search for helices and strands separately.
3. Use clusters to guide protein folding simulation.