

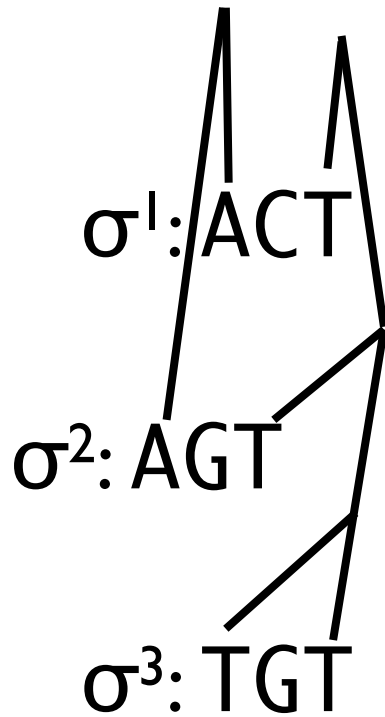
# BMI/CS 776

## Lecture #10:

# Alignment combinatorics

Colin Dewey  
2007.02.22

# Homology forests revisited



Set of positions in sequence  $i$  (forward strand)

$$S_{\sigma^i} = \{(i, j, +) : j \in \{1, \dots, n_i\}\}$$

Set of positions in all sequences (forward strand)

$$S_{\sigma^1, \dots, \sigma^k} = \bigcup_{i=1}^k S_{\sigma^i}$$

Set of positions in sequence  $i$  (reverse strand)

$$\bar{S}_{\sigma^i} = \{(i, j, -) : j \in \{1, \dots, n_i\}\}$$

Set of positions in all sequences (reverse strand)

$$\bar{S}_{\sigma^1, \dots, \sigma^k} = \bigcup_{i=1}^k \bar{S}_{\sigma^i}.$$

**homology forest:** forest with leaves labeled by  $S \cup \bar{S}$ ,  
with  $\#$  occurrences of label  $(i, j, +) + \#$  occurrences of label  $(i, j, -) \leq 1$

# Alignments today

- Just matching of homologous positions
- No trees, or trees are afterthought
- Homology sets = columns in alignment

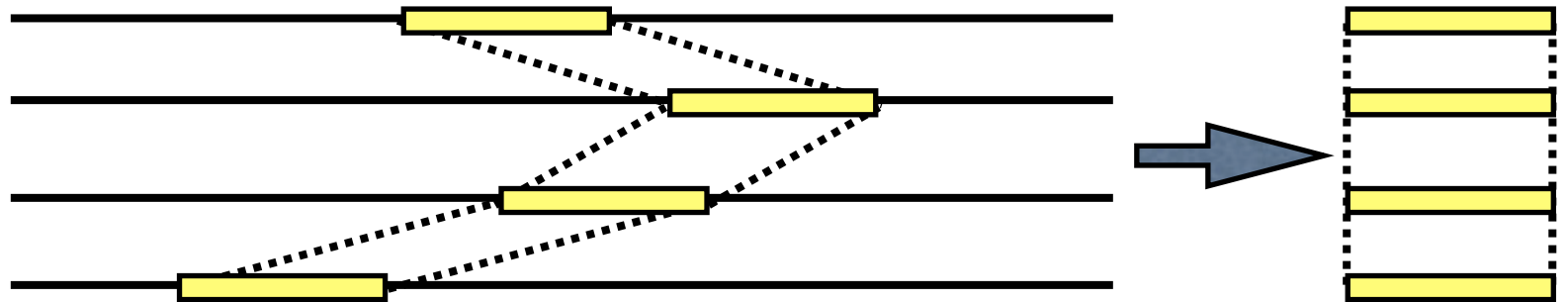
Diagram illustrating sequence alignment. A vertical yellow bar highlights a column of positions across five sequences, indicating they are homologous. An arrow points from the label 'homologous' to this column.

CA-GC--TACGGCTT	A-GCCT
TA-CCACTAC--CT	GA-GCAT
CA-GCAGTTC--CTT	A-GCCT
CA-GC--TACCGCT	GA-ACAT
CATGCAGTTC--CTT	ACACCT

homologous

# Alignments we've seen

- Motif finding
  - Local multiple alignment
  - May or may not be actually homologous
  - No trees

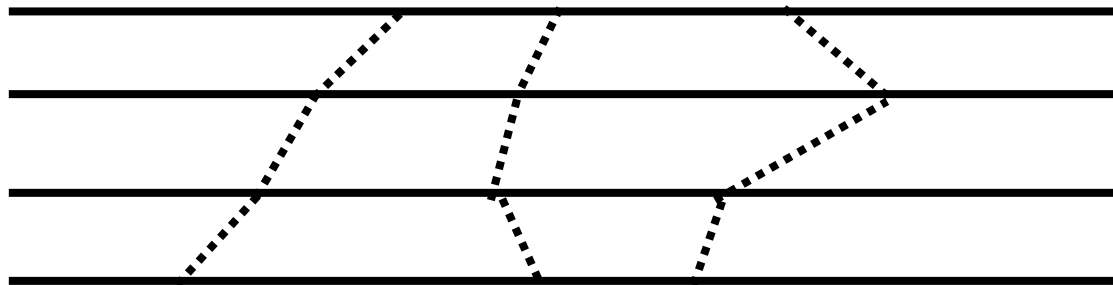


# Order in alignments

- Homology forests do not deal with order of nucleotides in either extant or ancestral species
- However, order is very important for
  - Determining homologous positions
  - Representing and visualizing homology

# Global alignment

- Given  $k$  homologous sequences
- Assumes sequences are all **colinear**:
- Homologous positions occur in the same order in each sequence



# Partially ordered set

- A **partially ordered set (poset)** is a set  $P$  together with a relation  $\leq$  with the following properties (for all  $x, y, z$  in  $P$ ):
  - reflexivity:  $x \leq x$
  - antisymmetry: If  $x \leq y$  and  $y \leq x$ , then  $x = y$
  - transitivity: If  $x \leq y$  and  $y \leq z$ , then  $x \leq z$

# Morning tasks poset

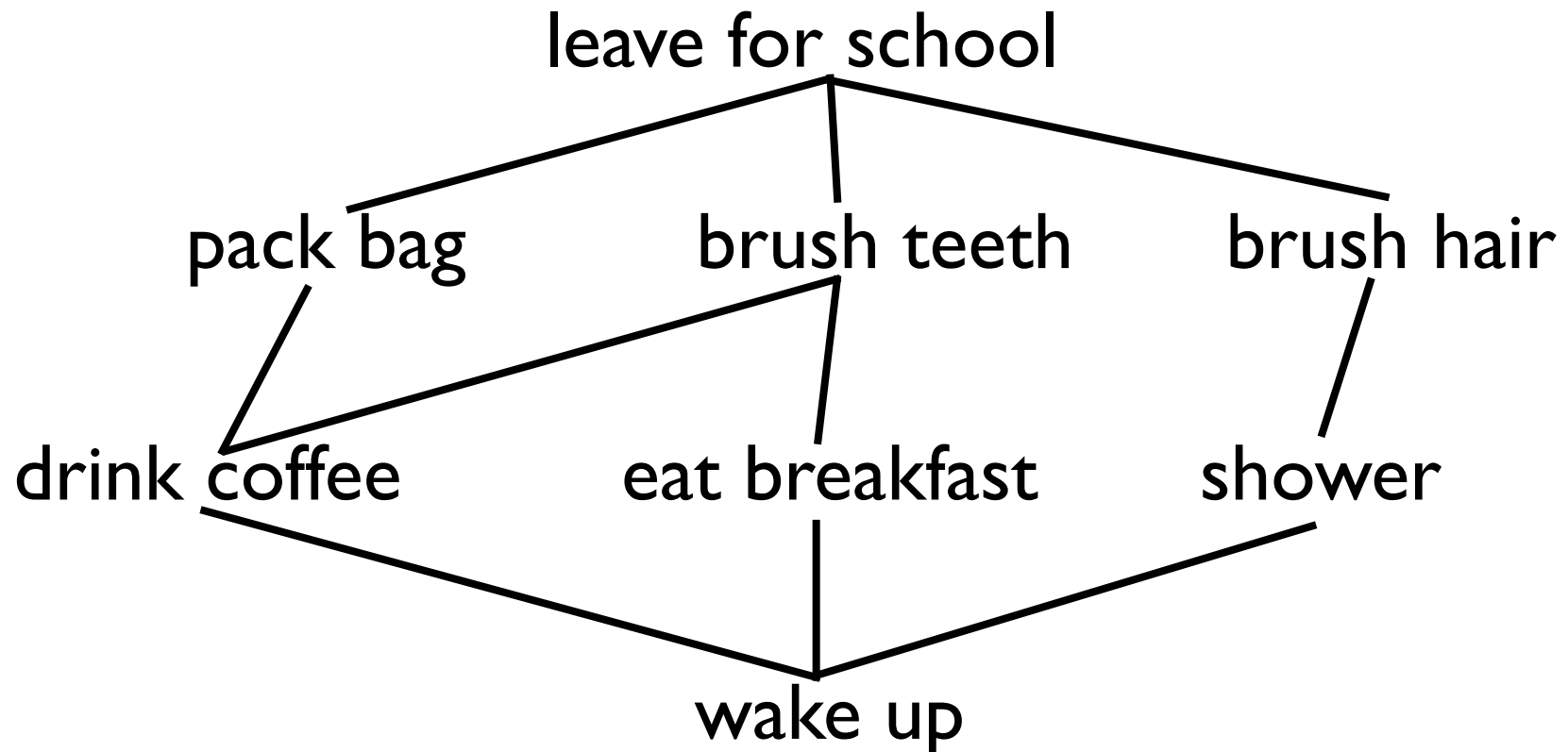
- $P = \{\text{wake up, brush hair, shower, leave for school, drink coffee, eat breakfast, pack bag, brush teeth}\}$
- $x \leq y$  if task  $y$  may not be done before task  $x$ 
  - shower  $\leq$  brush hair
  - eat breakfast  $\leq$  brush teeth
  - shower  $\leq$  leave for school
  - etc.





# Hasse diagram

- Line upward from  $x$  to  $y$  if  $x \leq y$  and there is no  $z \in P$  such that  $x \leq z \leq y$



# Partial global multiple alignment

A *partial global multiple alignment* of sequences  $\sigma^1, \dots, \sigma^k$  is a partially ordered set  $P = \{c_1, \dots, c_m\}$  together with a surjective function  $\varphi : S_{\sigma^1, \dots, \sigma^k} \rightarrow P$  such that  $\varphi((i, j_1, \epsilon_1)) \leq \varphi((i, j_2, \epsilon_2))$  if  $j_1 \leq j_2$ .

- $c_1, \dots, c_m$  : columns in multiple alignment
- $P$  : the “alignment poset”
- *surjective*:  $\varphi$  maps at least one sequence position to every column  $c_i$
- $(i, j, \epsilon)$ : position  $j$  in sequence  $i$  on strand  $\epsilon$

# Example alignment poset

$\varphi$

$(1,1,+) \rightarrow c_1$

$(1,2,+) \rightarrow c_2$

$(1,3,+) \rightarrow c_3$

$\sigma^1: \text{ACT}$

$(2,1,+) \rightarrow c_1$

$c_1 \leq c_2 \leq c_3$

$\sigma^2: \text{AGT}$

$(2,2,+) \rightarrow c_4$

$c_1 \leq c_4 \leq c_3$

$\sigma^3: \text{TGT}$

$(2,3,+) \rightarrow c_3$

$c_5 \leq c_4 \leq c_3$

$(3,1,+) \rightarrow c_5$

$(3,2,+) \rightarrow c_4$

$(3,3,+) \rightarrow c_3$

one possible alignment:

$c_5$	$c_1$	$c_4$	$c_2$	$c_3$
-	A	-	C	T
-	A	G	-	T
T	-	G	-	T

- Unless we have a **total order**, the order of some columns is not specified (e.g.,  $c_5$  and  $c_1$ )

# Extreme example: Null alignment

- Given sequences  $\sigma^1, \sigma^2, \dots, \sigma^k$  of lengths  $n_1, n_2, \dots, n_k$
- Null alignment: size of  $P$  is  $\sum n_i$ 
  - Every position is mapped to a different column

ACT-----      -----ACT      A---CT---  
 ---AGT---    or   AGT-----    or   -A-G--T--    or...  
 -----TGT      ---TGT---      --T----GT

# Number of pairwise partial global alignments

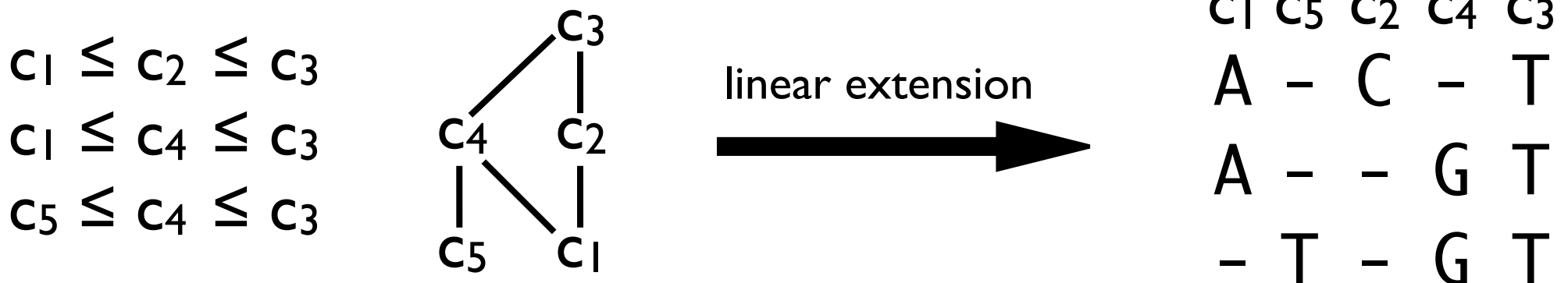
**Proposition 2.5** *The number of partial global alignments of two sequences of length  $n$  and  $m$  is  $\binom{n+m}{m}$ .*

**Proof:** Note that the number of alignments with  $k$  homologous nucleotides is given by  $\binom{n}{k} \binom{m}{k}$ . The total number of alignments is therefore

$$\sum_{k=0}^{\min(n,m)} \binom{n}{k} \binom{m}{k} = \binom{n+m}{n}.$$

# Linear extension

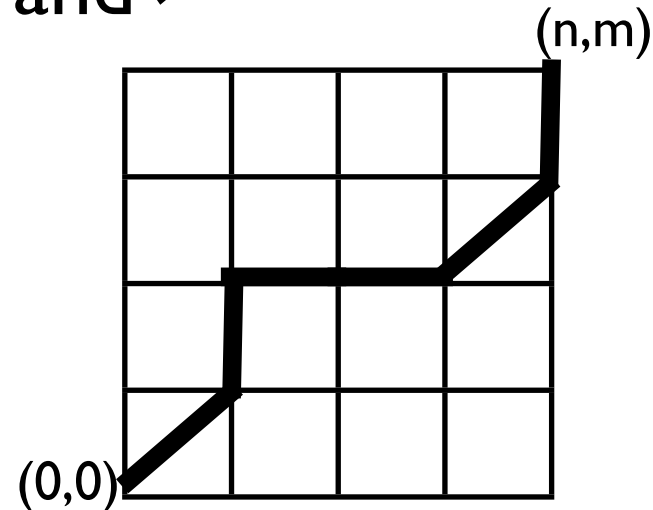
- A **linear extension** of a partially ordered set  $P = \{c_1, \dots, c_m\}$  is a permutation  $\pi$  of the elements  $c_1, \dots, c_m$  such that whenever  $c_i \leq c_j$ ,  $\pi(c_i) \leq \pi(c_j)$
- A **global multiple alignment** is a partial global multiple alignment along with a linear extension of the alignment poset



# Number of pairwise global alignments

- The number of pairwise global alignments of sequences of length  $n$  and  $m$  is the Delannoy number  $D_{n,m}$
- $D_{n,m}$ : number of lattice paths from  $(0,0)$  to  $(n,m)$  with allowed moves  $\uparrow$ ,  $\rightarrow$ , and  $\nearrow$

$$D_{n,m} = \sum_{k=0}^{\min(n,m)} \binom{n}{k} \binom{m}{k} 2^k$$



# Partial vs full alignments

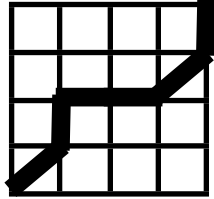
- The number of possible partial alignments is very large
- The number of full alignments is even larger

Number of pairwise alignments

n	partial	full
1	2	3
2	6	13
3	20	63
4	70	321
5	252	1,683
6	924	8,989
7	3,432	48,639
8	12,870	265,729
9	48,620	1,462,563
10	184,756	8,097,453



# Representing full pairwise global alignments

- representations for pairwise alignment  $h$ , of sequences  $\sigma^1 = \sigma_1^1 \sigma_2^1 \cdots \sigma_n^1$  and  $\sigma^2 = \sigma_1^2 \sigma_2^2 \cdots \sigma_m^2$
- edit string  $h$  over edit alphabet  $\{H, I, D\}$ 
  - H: homology, I: insertion, D: deletion
- path in alignment graph 
- sequence of pairs  $(\sigma_i^1 \diamond \sigma_j^2)$ ,  $(\sigma_i^1 \diamond -)$ , or  $(\sigma_i^2 \diamond -)$

# Comparing alignments

- Compare in terms of H, I, and D pairs:

$$h_H = \{(i, j) : (\sigma_i^1 \diamond \sigma_j^2) \in h\},$$

$$h_D = \{i : (\sigma_i^1 \diamond -) \in h\},$$

$$h_I = \{j : (\sigma_j^2 \diamond -) \in h\}.$$

for any  $h \in \mathcal{A}_{n,m}$        $|h_H| + |h_D| = n$  and  $|h_H| + |h_I| = m$ .

# Alignment equivalence

- Alignments are defined to be equivalent if they match up the same sequence positions

$$h^i \sim h^j \text{ if and only if } h_H^i = h_H^j$$

- Equivalently, alignments are equivalent if they gap the same sequence positions

$$h^i \sim h^j \text{ if and only if } h_I^i = h_I^j \text{ and } h_D^i = h_D^j$$

$$\begin{array}{ccc} \text{AC-T} & & \text{A-CT} \\ \text{A-GT} & \sim & \text{AG-T} \\ h^i & & h^j \end{array}$$

# Alignment measures

- Measures of sensitivity and specificity, with respect to a reference alignment  $h^r$

$$f(h^i, h^j) = \frac{|h_H^i \cap h_H^j|}{|h_H^i|}$$

$$f_D(h^p, h^r) = f(h^r, h^p) = \frac{|h_H^r \cap h_H^p|}{|h_H^r|}$$

“developer’s measure”  
(sensitivity)

$$f_M(h^p, h^r) = f(h^p, h^r) = \frac{|h_H^r \cap h_H^p|}{|h_H^p|}$$

“modeler’s measure”  
(specificity)

# Distances between alignments

- Would like a distance function between alignments
- Should be a metric, i.e., it should satisfy:

$$d(h^i, h^j) \geq 0$$

$$\forall h^i, h^j \in \mathcal{A}_{n,m},$$

$$d(h^i, h^j) = 0 \text{ iff } h^i \sim h^j$$

$$\forall h^i, h^j \in \mathcal{A}_{n,m},$$

$$d(h^i, h^j) = d(h^j, h^i)$$

$$\forall h^i, h^j \in \mathcal{A}_{n,m},$$

$$d(h^i, h^j) + d(h^j, h^k) \geq d(h^i, h^k)$$

$$\forall h^i, h^j, h^k \in \mathcal{A}_{n,m}.$$

# Alignment metric

- The following function is a finite metric on alignments:

$$\begin{aligned}d(h^i, h^j) &= 2|h_H^i| + |h_I^i| + |h_D^i| - 2|h_H^i \cap h_H^j| \\&\quad - |h_I^i \cap h_I^j| - |h_D^i \cap h_D^j| \\&= 2|h_H^j| + |h_I^j| + |h_D^j| - 2|h_H^i \cap h_H^j| \\&\quad - |h_I^i \cap h_I^j| - |h_D^i \cap h_D^j| \\&= n + m - 2|h_H^i \cap h_H^j| \\&\quad - |h_I^i \cap h_I^j| - |h_D^i \cap h_D^j|).\end{aligned}$$

(Schwartz & Pachter, 2007)

# Example alignment metric

Metric for  $\mathcal{A}_{2,2}$

	<i>HH</i>	<i>HDI</i>	<i>DIH</i>	<i>IHD</i>	<i>DHI</i>	<i>DDII</i>
<i>HH</i>	0	2	2	4	4	4
<i>HDI</i>	2	0	4	3	3	2
<i>DIH</i>	2	4	0	3	3	2
<i>IHD</i>	4	3	3	0	4	2
<i>DHI</i>	4	3	3	4	0	2
<i>DDII</i>	4	2	2	2	2	0

$$d(h^i, h^j) = n + m - 2|h_H^i \cap h_H^j| - |h_I^i \cap h_I^j| - |h_D^i \cap h_D^j|.$$

# Alignment metric accuracy

- Instead of developer or modeler score, use score based on metric

$$g(h^i, h^j) = 1 - \frac{d(h^i, h^j)}{n+m}$$

- $g$  is fraction of positions aligned identically in the two alignments
- Alignment Metric Accuracy (AMA) =  $g(h^p, h^r)$

(Schwartz & Pachter, 2007)



# Multiple alignment accuracy

- All multiple alignments of sequences of lengths  $n_1, n_2, \dots, n_k$ :  $\mathcal{A}_{n_1, n_2, \dots, n_k}$

Given two MSAs  $h^i, h^j \in \mathcal{A}_{n_1, n_2, \dots, n_k}$ :

$$d(h^i, h^j) = \sum_{s^1=1}^{k-1} \sum_{s^2 > s^1}^k d(h_{s^1, s^2}^i, h_{s^1, s^2}^j)$$

- Like sum-of-pairs scoring
- Accuracy:  $g(h^p, h^r) = 1 - \frac{d(h^p, h^r)}{(k-1) \sum_{i=1}^k n_i}$ .

(Schwartz & Pachter, 2007)

# Next time

- Statistical models for pairwise alignment
- Evolutionary models for sequences undergoing insertions and deletions
- Algorithms for Alignment Metric Accuracy