Scores and substitution matrices in inexact matching (sequence alignment)

CMSC423

Slides courtesy of Sushmita Roy
U. Wisconsin

Key concepts in this unit

- Probabilistic interpretation of scores in alignment algorithms
- Different substitution matrices
- Assessing significance of scores
  - Bayesian approach
  - Extreme Value Theory
- Heuristic algorithms to speed up search (BLAST)

Readings

- Sections 2.1, 2.2
- Sections 2.3 till end of Smith-Waterman
- Section 2.7
- Section 2.8

Guiding principles of scores in alignments

- Sequence is said to have diverged from a common ancestor through mutations
  - Substitutions
  - Insertions and deletions (gaps)
- Score evolutionarily close alignments higher than those that are not
- That is we compute the **likelihood ratio** of an alignment given the two sequences are related versus not related

PROBABILITY PRIMER

Sample spaces

- *Sample space*: a set of possible outcomes for some event
- examples
  - flight to Chicago: {on time, late}
  - lottery: {ticket 1 wins, ticket 2 wins,…,ticket *n* wins}
  - weather tomorrow:

    {rain, not rain} or

    {sun, rain, snow} or

    {sun, clouds, rain, snow, sleet} or…

Random variables

- *Random variable*: represents the outcome of an experiment

- Example
  - X represents the outcome of my flight to Chicago
  - we write the probability of my flight being on time as $P(X = \text{on-time})$
  - or when it's clear which variable we're referring to, we may use the shorthand $P(\text{on-time})$

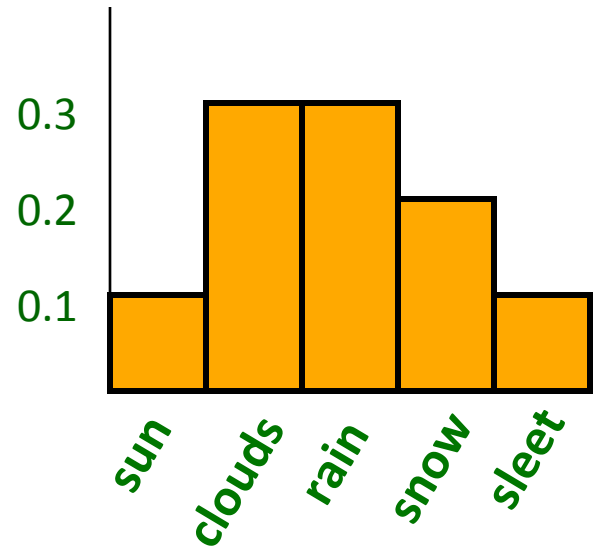
Probability distributions

- If X is a random variable, the function given by $P(X = x)$ for each x is the *probability distribution* of X

- Requirements:

$$P(x) \geq 0 \quad \text{for every } x$$

$$\sum_x P(x) = 1$$



Joint distributions

- *Joint probability distribution*: the function given by $P(X = x, Y = y)$
- Read "X equals x <u>and</u> Y equals y"
- Example

x, y	$P(X = x, Y = y)$
sun, on-time	0.20
rain, on-time	0.20
snow, on-time	0.05
sun, late	0.10
rain, late	0.30
snow, late	0.15

probability that it's sunny and my flight is on time

Marginal distributions

- The *marginal distribution* of X is defined by

$$P(x) = \sum_y P(x,y)$$

"the distribution of X ignoring other variables"

- This definition generalizes to more than two variables, e.g.

$$P(x) = \sum_y \sum_z P(x,y,z)$$

Marginal distribution example

joint distribution

x, y	$P(X = x, Y = y)$
sun, on-time	0.20
rain, on-time	0.20
snow, on-time	0.05
sun, late	0.10
rain, late	0.30
snow, late	0.15

marginal distribution for X

x	$P(X = x)$
sun	0.3
rain	0.5
snow	0.2

Conditional distributions

- The *conditional distribution* of X given Y is defined as:

$$P(X = x \mid Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)}$$

"the distribution of X given that we know the value of Y"

Conditional distribution example

joint distribution

x, y	$P(X = x, Y = y)$
sun, on-time	0.20
rain, on-time	0.20
snow, on-time	0.05
sun, late	0.10
rain, late	0.30
snow, late	0.15

conditional distribution for X given Y=on-time

x	$P(X = x \mid Y = on\text{-}time)$
sun	0.20/0.45 = 0.444
rain	0.20/0.45 = 0.444
snow	0.05/0.45 = 0.111

Independence

- Two random variables, X and Y, are *independent* if

$$P(x,y) = P(x) \times P(y) \quad \text{for all } x \text{ and } y$$

Independence example #1

joint distribution

x, y	$P(X = x, Y = y)$
sun, on-time	0.20
rain, on-time	0.20
snow, on-time	0.05
sun, late	0.10
rain, late	0.30
snow, late	0.15

marginal distributions

x	$P(X = x)$
sun	0.3
rain	0.5
snow	0.2

y	$P(Y = y)$
on-time	0.45
late	0.55

Are X and Y independent here? NO.

Independence example #2

joint distribution

x, y	$P(X = x, Y = y)$
sun, fly-United	0.27
rain, fly-United	0.45
snow, fly-United	0.18
sun, fly-Northwest	0.03
rain, fly-Northwest	0.05
snow, fly-Northwest	0.02

marginal distributions

x	$P(X = x)$
sun	0.3
rain	0.5
snow	0.2

y	$P(Y = y)$
fly-United	0.9
fly-Northwest	0.1

Are X and Y independent here? YES.

Log odds score

- Let X be a random variable representing an alignment
- Let M_1 and M_2 be two probabilistic models for X
- Log odds score $S(X)$

$$S(X) = \log \frac{P(X|M_1)}{P(X|M_2)}$$

- If $S(X)>0$, X is more likely to come from model M_1
- If $S(X)<0$, X is more likely to come from model M_2

What are M_1 and M_2 in our sequence alignment problem

- M_1: foreground model, that is the sequences are "related by evolution".
- M_2: background model, that is the sequences are unrelated
- Need to compute the probability of an alignment X, under the two models M_1 and M_2
- Assume alignments on **protein sequences** with no gaps.

M_1: foreground model

- Assume each pair of aligned positions evolved from a common ancestor
- Let p_{ab} be the probability of observing a pair $\{a,b\}$
- Probability of an alignment between x and y is

$$P(x, y|M_1) = \prod_{i=1}^{n} p_{x_i y_i}$$

M_2: background model

- Assume the individual amino acids at a position are independent of the amino acid in another position.
- Let q_a be the probability of amino acid a
- The probability of an n-character alignment of x and y is

$$P(x, y|M_2) = \prod_{i=1}^{n} q_{x_i} \prod_{i=1}^{n} q_{y_i}$$

Computing the log odds ratio to score an alignment

- The score of an alignment is the log odds ratio of the two sequences from M_1 and M_2

$$S = \log \frac{P(x, y|M_1)}{P(x, y|M_2)}$$

$$S = \log \frac{\prod_{i=1}^{n} p_{x_i y_i}}{\prod_{i=1}^{n} q_{x_i} q_{y_i}}$$

Computing the log odds ratio to score an alignment

$$S = \sum_{i=1}^{n} \log \frac{p_{x_i y_i}}{q_{x_i} q_{y_i}}$$

Score of an alignment

$$s(a, b) = \log \frac{p_{a,b}}{q_a q_b}$$

Substitution matrix entry

Some common substitution matrices

- BLOSUM matrices [Henikoff and Henikoff, 1992]
  - BLOSUM45
  - BLOSUM50
  - BLOSUM62
    - Number represents percent identity of sequences used to construct substitution matrices
- PAM [Dayhoff et al, 1978]
- Empirically, BLOSUM62 works the best

How to estimate the probabilities?

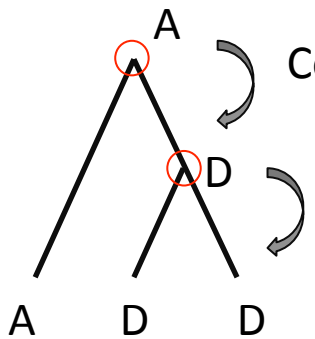
- Need a good set of confirmed alignments
- Depends upon what we know about when the two sequences might have diverged
  - p_{ab} for closely related species is likely to be low if $a\ !=b$
  - p_{ab} for species that have diverged a long time ago is likely close to the background.

Dayhoff Point accepted mutation (PAM) matrix

- Substitution data from very similar/evolutionary close proteins
  - 71 protein sequences
- Estimate ancestral sequence based on parsimony
  - We will look at this in detail in Phylogenetic trees
- Estimate A_{ab} the frequency of observing a,b pair in ancestor child pairs.
- Derive a conditional probability of $P(a|b)$ for unit time.
- Derive a condition probability for longer time by taking powers of the conditional probability matrix.

Calculating Dayhoff PAM matrices

- Ancestral points



Count number of amino acid pairs for each ancestor – child pair

A_{ab} Total number of observed a,b pairs

$$P(a|b) = \frac{A_{ab}}{\sum_c A_{bc}}$$

Conditional probability in unit time

BLOSUM matrices

- BLOck Substitution Matrix
- Derived from a set of aligned ungapped regions from protein families called BLOCKS
- Cluster proteins such that they have no less than L% of similarity

Different BLOSUM matrices

- BLOSUM50
  - Proteins >50% similarity are in the same group
- BLOSUM62
  - Proteins >62% similarity are in the same group

Example substitution scoring matrix (BLOSUM62)

BLOSUM62

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	X
A	4																				
R	-1	5																			
N	-2	0	6																		
D	-2	-2	1	6																	
C	0	-3	-3	-3	9																
Q	-1	1	0	0	-3	5															
E	-1	0	0	2	-4	2	5														
G	0	-2	0	-1	-3	-2	-2	6													
H	-2	0	1	-1	-3	0	0	-2	8												
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4											
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4										
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5									
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5								
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6							
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7						
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4					
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5				
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11			
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7		
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4	
X	0	-1	-1	-1	-2	-1	-1	-1	-1	-1	-1	-1	-1	-1	-2	0	0	-2	-1	-1	-1

Positive for chemically similar substitution

Common amino acids have low weights

Rare amino acids have high weights

Conserved blocks

```
AABCDA...BBCDA
DABCDA.A.BBCBB
BBBCDABA.BCCAA
AAACDAC.DCBCDB
CCBADAB.DBBDCC
AAACAA...BBCCC
```

Block1 Block2

From Ron Shamir's lecture http://www.cs.tau.ac.il/~rshamir/algmb/98/scribe/html/lec03/node10.html

Estimating the probabilities in BLOSUM

$$p_{ab} = \frac{A_{ab}}{\sum_{cd} A_{cd}}$$

$$q_a = \frac{\sum_b A_{ab}}{\sum_{cd} A_{cd}}$$

$$s(a,b) = \log \frac{p_{a,b}}{q_a q_b}$$

Calculating the probabilities

$A_{ab}^{(k)}$: Number of ab pairs in the kth column of a block

```
AABCDA...BBCDA
DABCDA.A.BBCBB
BBBCDABA.BCCAA
AAACDAC.DCBCDB
CCBADAB.DBBDCC
AAACAA...BBCCC
```

$$A_{ab} = \sum_k A_{ab}^{(k)}$$

$$\mathrm{A}_{AA}^{(1)} = 6 \qquad A_{AB}^{(1)} = 4$$

$$A_{AC}^{(1)} = 4 \qquad A_{BC}^{(1)} = 1$$

Estimating significance of scores

- How do we know whether a given alignment score is random or significant?
- Two approaches
  - Bayesian Approach
  - A classical approach: the extreme value distribution

Bayesian approach

- Recall in our log odds ratio we estimated

$$P(x, y|\mathbf{M_1})$$ Related

$$P(x, y|\mathbf{M_2})$$ Unrelated

- We could instead ask what is the probability of the two sequences being related as opposed to unrelated

$$P(\mathbf{M_1}|x, y)$$

Bayes theorem

$$P(x \mid y) = \frac{P(y \mid x)P(x)}{P(y)} = \frac{P(y \mid x)P(x)}{\sum_x P(y \mid x)P(x)}$$

- An extremely useful theorem
- There are many cases when it is hard to estimate $P(x \mid \mathrm{y})$ directly, but it's not too hard to estimate $P(y \mid x)$ and $P(x)$

Bayes theorem example

- MDs usually aren't good at estimating $P(Disorder | Symptom)$
- They're usually better at estimating $P(Symptom \mid Disorder)$
- If we can estimate $P(Fever \mid Flu)$ and $P(Flu)$ we can use Bayes' Theorem to do diagnosis

$$P(flu \mid fever) = \frac{P(fever \mid flu)P(flu)}{P(fever \mid flu)P(flu) + P(fever \mid \neg flu)P(\neg flu)}$$

Using Bayes Rule to estimate $P(M_1|x,y)$

$$P(\mathbf{M_1}|x,y) = \frac{P(x,y|\mathbf{M_1})P(\mathbf{M_1})}{P(x,y)}$$

Bayes rule

$$P(\mathbf{M_1}|x,y) = \frac{P(x,y|\mathbf{M_1})P(\mathbf{M_1})}{P(x,y,\mathbf{M_1}) + P(x,y,\mathbf{M_2})}$$

Marginalization

Chain Rule

$$P(\mathbf{M_1}|x,y) = \frac{P(x,y|\mathbf{M_1})P(\mathbf{M_1})}{P(x,y|\mathbf{M_1})P(\mathbf{M_1}) + P(x,y|\mathbf{M_2})P(\mathbf{M_2})}$$

Using Bayes Rule to estimate $P(M_1|x,y)$

$$P(\mathbf{M_1}|x,y) = \frac{P(x,y|\mathbf{M_1})P(\mathbf{M_1})}{P(x,y)}$$

Bayes rule

$$P(\mathbf{M_1}|x,y) = \frac{P(x,y|\mathbf{M_1})P(\mathbf{M_1})}{P(x,y,\mathbf{M_1}) + P(x,y,\mathbf{M_2})}$$

Marginalization

Chain Rule

$$P(\mathbf{M_1}|x,y) = \frac{P(x,y|\mathbf{M_1})P(\mathbf{M_1})}{P(x,y|\mathbf{M_1})P(\mathbf{M_1}) + P(x,y|\mathbf{M_2})P(\mathbf{M_2})}$$

Model priors

Points about $P(M_1|x,y)$

- Has the form of a logistic function

$$P(\mathbf{M_1}|x,y) = \frac{P(x,y|\mathbf{M_1})P(\mathbf{M_1})/P(x,y|\mathbf{M_2})P(\mathbf{M_2})}{1 + P(x,y|\mathbf{M_1})P(\mathbf{M_1})/P(x,y|\mathbf{M_2})P(\mathbf{M_2})}$$

$$P(\mathbf{M_1}|x,y) = \frac{e^x}{1+e^x}$$

where

$$x = \log P(x,y|\mathbf{M_1})P(\mathbf{M_1})/P(x,y|\mathbf{M_2})P(\mathbf{M_2})$$

$$= \underbrace{\log\frac{P(x,y|\mathbf{M_1})}{P(x,y|\mathbf{M_2})}}_{\text{Alignment score}} + \underbrace{\log\frac{P(\mathbf{M_1})}{P(\mathbf{M_2})}}_{\text{Prior log odds score}}$$

Points about $P(\boldsymbol{M_1}|x,y)$

- The prior log odds score is added to the sequence score
- This can be used to encode our prior belief of expected number of matches
- In fact the prior log odds score should be inversely related to the number of sequences we have in a database

The classical approach to assessing sequence :Extreme Value Distribution

- Suppose we have a particular substitution matrix and amino-acid frequencies
- We need to consider random sequences of lengths m and n and finding the best alignment of these sequences
- This will give us a distribution over alignment scores for random pairs of sequences
- If the probability of a random score being greater than our alignment score is small, we can consider our score to be significant

Scores from random alignments

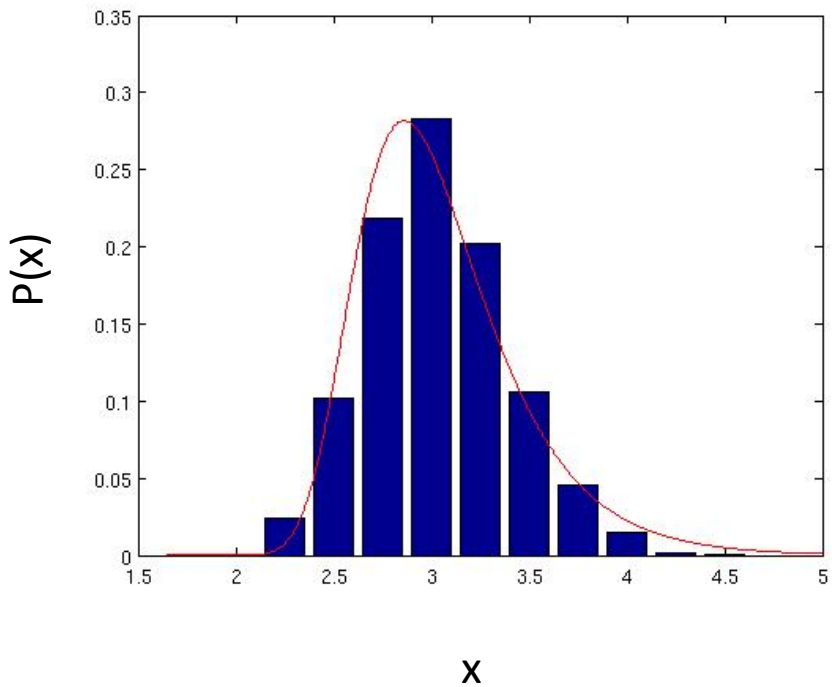
- Suppose we assume
  - Sequence lengths m and n
  - A particular substitution matrix and amino-acid frequencies

- And we consider generating random sequences of lengths m and n and finding the best alignment of these sequences

- This will give us a distribution over alignment scores for random pairs of sequences

The extreme value distribution

- Because we're picking the best alignments, we want to know the distribution of max scores for alignments against a random set of sequences looks like
- this is given by an *extreme value distribution*



Assessing significance of sequence score alignments

- It can be shown that the mode of the distribution for optimal scores is

$$U = \frac{\log(Kmn)}{\lambda}$$

  - K, λ estimated from the substitution matrix
- Probability of observing a score greater than S

$$P(x > S) = 1 - \exp(-\exp^{-\lambda(S-U)})$$

$$P(x > S) = 1 - \exp(-Kmn\exp^{-\lambda S)})$$

Need to speed up sequence alignment

- consider the task of searching the RefSeq collection of sequences against a query sequence:
  - most recent release of DB contains 32,504,738 proteins
  - Entails 33,000,000*(300*300) matrix operations (assuming query sequence is of length 300 and avg. sequence length is 300)
- O(mn) too slow for large databases with high query traffic

Speeding up sequence alignment

- Indexing techniques to locate possible small high scoring segments
- Throw away segments that are not significant (based on theory of score significance)
- Extending only high scoring segments
- Two heuristic algorithms
  - BLAST
  - FASTA

BLAST: Basic Local Alignment Search Tool

- Altshul et al 1990
  - Cited >48,000 times!
- Optimizes Maximal Segment Pair (MSP) score
  - A local measure of similarity
- Used EVD like theory for random sequence score
- Works for both protein sequence and DNA sequence
  - Only scores differ

Maximal Segment Pair (MSP)

- Sequence segment: A contiguous stretch of residues of any length
- **Relies on key assumption of addivity:**
  - Similarity score for two aligned segments of the same length is the sum of similarity values for each pair of aligned residues.
- MSP: highest scoring pair of identical length segments from two sequences
- Theoretical analysis gives the statistical significance of an MSP score
  - Allows BLAST to efficiently prune out low scoring pairs

BLAST continued

- BLAST finds locally maximal segment pairs that exceeds a particular cutoff
- Let a word pair be a segment pair of length w
- BLAST only seeks those word pairs that have a score at least T
- Extend only word pairs with a score of at least T to determine if it has a segment pair of score at least S.

Key steps of the BLAST algorithm

- For each query sequence
  1. Compile a list of high-scoring words of score at least T
     - First generate words in the query sequence
     - Then find words that match query sequence words with score at least T
     - Thus allows for inexact matches
  2. Scan the database for hits of these words
  3. Extend hits

Determining query words

Given:

query sequence: QLNFSAGW

word length $w = 2$ (default for protein usually $w = 3$)

word score threshold $T = 9$

Step 1: determine all words of length w in query sequence

QL LN NF FS SA AG GW

Determining query words

Determine all words that score at least *T* when compared to a word in the query sequence

words from query sequence	words with T≥9
QL	QL=9
LN	LN=10
NF	NF=12, NY=9
…	
SA	none
…	

Additional words in the database

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	4																			
R	-1	5																		
N	-2	0	6																	
D	-2	-2	1	6																
C	0	-3	-3	-3	9															
Q	-1	1	0	0	-3	5														
E	-1	0	0	2	-4	2	5													
G	0	-2	0	-1	-3	-2	-2	6												
H	-2	0	1	-1	-3	0	0	-2	8											
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4										
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4									
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5								
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5							
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6						
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7					
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4				
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5			
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11		
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

Scanning the database

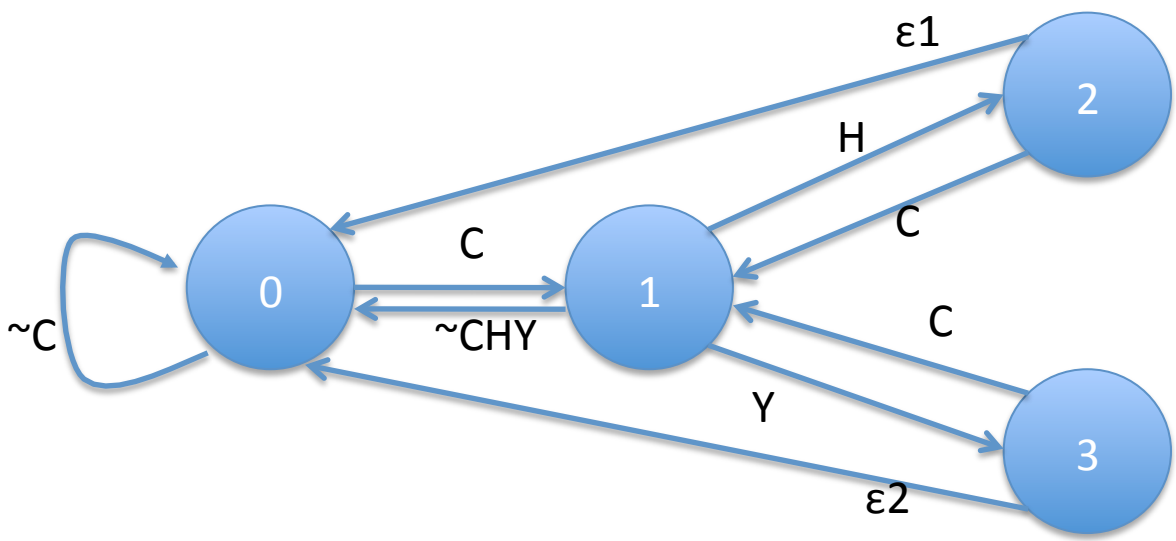
- How to efficiently search a long sequence for multiple occurrences of short sequences
- BLAST has two approaches
  - Indexing approach
  - Finite state machine

Indexing approach

- Let w=3. For amino acids, the number of words is 20^3.
- Map a word to an integer between 1 and 20^3.
- Thus a word has an index into an array
- Each index points to a list of matches of the word in the query sequence
- As we scan the database, each database word immediately leads to the hits in the query sequence

Deterministic finite state machine (FSM)

- Deterministic behavior as input is read
  - State transitions/outputs
- An example FSM to match CHY, CHH and CYH



Transition from 2 or 3 to 0, happens when C is not observed. Can result in “restarting” or “acceptance”

Extending a hit

- Extending a word hit to a segment pair is straightforward
- Terminate extension when the score of the pair falls a certain distance below the best score found for shorter extensions

How to choose w and T?

- Tradeoff between running time and sensitivity
- Sensitivity

$$\text{sensitivity} = \frac{\#\text{ significant matches found}}{\#\text{ of significant matches in DB}}$$

- T
  - small *T*: greater sensitivity, more hits to expand
  - large *T*: lower sensitivity, fewer hits to expand
- W
  - Larger w: fewer query word seeds, lower time for extending, but more possible words (20^w for AAs)
- In practice w=4, T=17 is good for proteins

Summary of BLAST

- T: Don’t consider seeds with score < T
- Don’t extend hits when score falls below a specified threshold
- Pre-processing of database or query helps to improve the running time

FASTA

- Starts with exact seed matches instead of inexact matches that satisfy a threshold
- Extends seeds (similar to BLAST)
- Join high scoring seeds allowing for gaps
- Re-align high scoring matches

Different versions of BLAST programs

Program	Query	Database
BLASTP	Protein	Protein
BLASTN	DNA	DNA
BLASTX	Translated DNA	Protein
TBLASTN	Protein	Translated DNA
TBLASTX	Translated DNA	Translated DNA

Sequence databases

- Large database centers
  - NCBI: http://www.ncbi.nih.gov
  - EBI: http://www.ebi.ac.uk
  - Sanger: http://www.sanger.ac.uk
  - Each of these centers link to hundreds of databases
- Nucleotide sequences
  - Genbank
  - EMBL-EBI Nucleotide Sequence Database
  - Comprise ~8% of the total database (Nucleic Acid Research 2006 Database edition)
- Protein sequences
  - UniProtKB

Using BLAST

- http://blast.ncbi.nlm.nih.gov/Blast.cgi

- Will blast a DNA sequence against NCBI nucleotide database
- We will select
  - http://www.ncbi.nlm.nih.gov/nuccore/NG_000007.3?from=70545&to=72150&report=fasta