Error Correction
30 Sept 15

Announcements

Final Project

- Groups of 3-5. No more than 1 grad per group.

- Can use data from published manuscript, downloaded from ENA, your own data.

- I approve topic (Oct14)

- Presentations last week of class

- Written paper due last day.

Final Project Examples

Genome

Transcriptome

Expression

Comparative Genomics

Exam review

type	query	database
blastn	nt	nt
blastx	Trans(nt)	prot
tblastx	Trans(nt)	Trans(nt)
blastp	prot	prot
tblastn	prot	Trans(nt)

BLAST

1. Build Lookup table

Preprocess: Build a *lookup table* of size $|\Sigma|^w$ for all w-length words in D

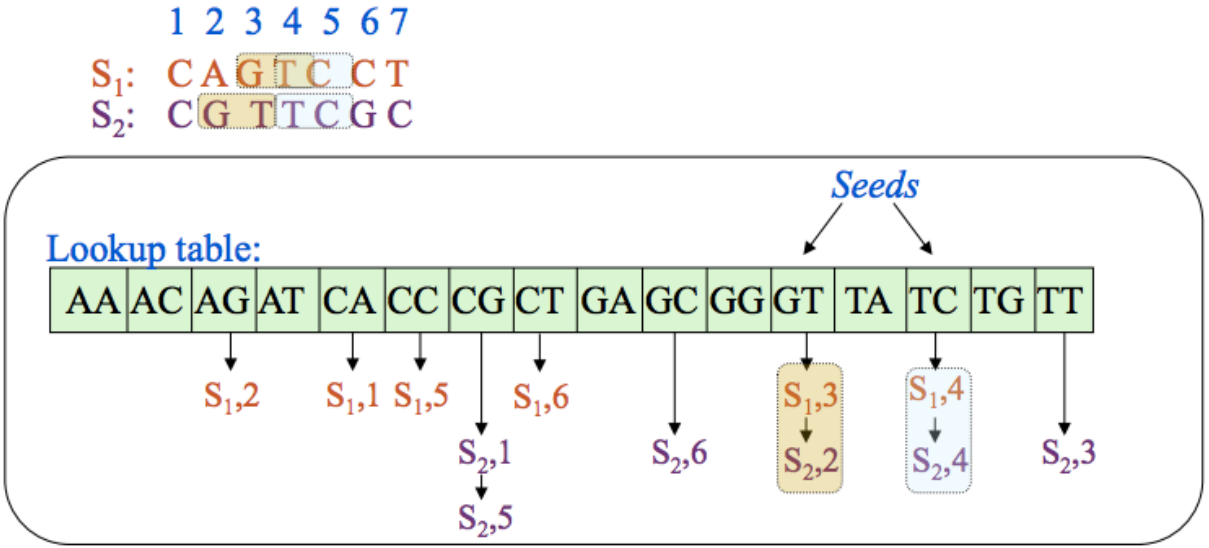
Σ={A,C,G,T}
$w = 2$
→ 4^2 (=16) entries in lookup table

Lookup table:

AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT

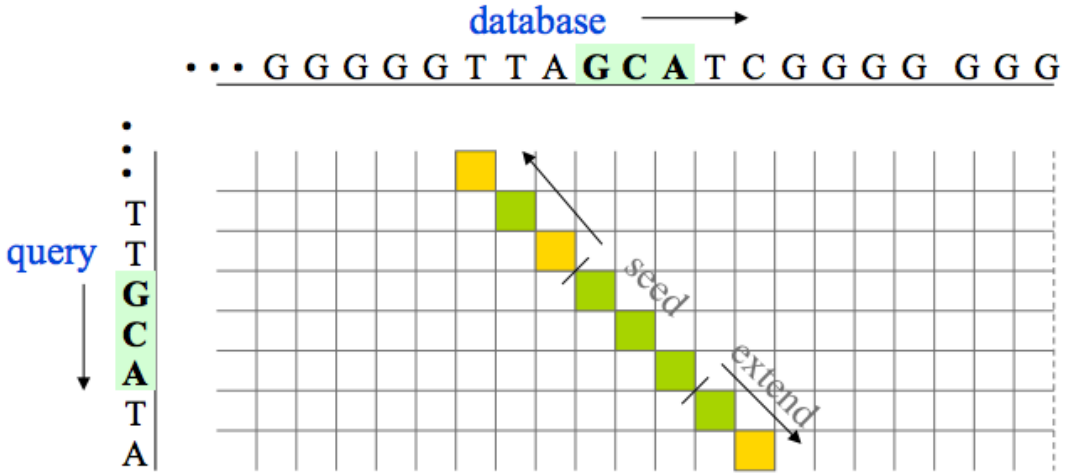
BLAST

2. Filter low complexity and Identify Seeds



BLAST

3. Bidirectional extension – (Smith Waterman algorithm)



BLAST

Stats

$$E = Kmne^{-\lambda S}$$

BLAST

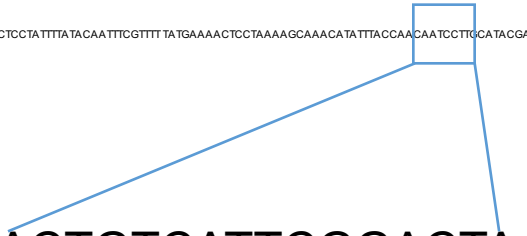
Stats

$$p = 1 - e^{-E}$$

Error Correction

Error Correction

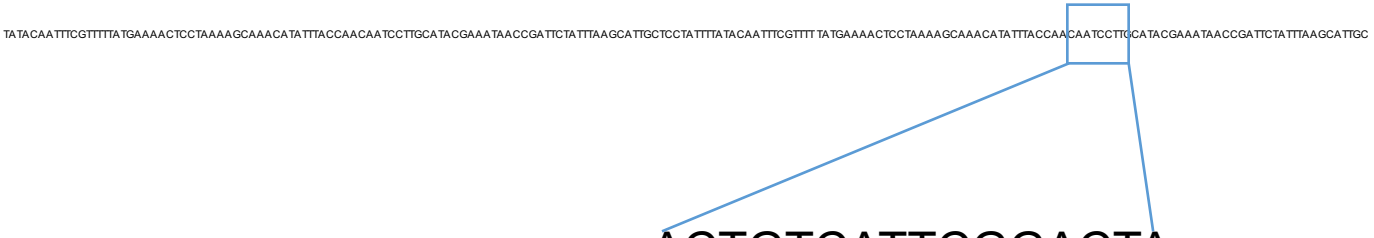
TATACAATTTCGTTTTTATGAAAACTCCTAAAAGCAAACATATTTACCAACAATCCTTGCATACGAAATAACCGATTCTATTTAAGCATTGCTCCTATTTTATACAATTTCGTTTTTATGAAAACTCCTAAAAGCAAACATATTTACCAACAATCCTTGCATACGAAATAACCGATTCTATTTAAGCATTGC



ACTGTCATTCGGACTA
ACTGTCATTCGGACTA
ACTGTCATTCGGCCTA
ACTGTCATTCGGACTA
ACTGTCATTCGGACTA
ACTGTCATTCGGACTA
ACTGTCATTCGGACTA
ACTGACATTCGGACTA
ACTGTCATTCGGACTA
ACTGTCATTCGGACTA

Consensus= ACTGTCATTCGGACTA

Error Correction

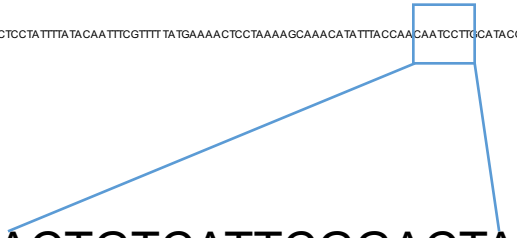


ACTGTCATTCGGACTA
ACTGTCATTCGGACTA
ACTGTCATTCGGCCTA
ACTGTCATTCGGACTA
ACTGTCATTCGGACTA
ACTGTCATTCGGACTA
ACTGTCATTCGGACTA
ACTGACATTCGGACTA
ACTGACATTCGGACTA
ACTGTCATTCGGACTA

Consensus= ACTG{A,T}CATTCGGACTA

Error Correction

TATACAATTTCGTTTTTATGAAAACTCCTAAAAGCAAACATATTTACCAACAATCCTTGCATACGAAATAACCGATTCTATTTAAGCATTGCTCCTATTTTATACAATTTCGTTTT TATGAAAACTCCTAAAAGCAAACATATTTACCAACAATCCTTGCATACGAAATAACCGATTCTATTTAAGCATTGC



ACTGTCATTCGGACTA
ACTGTCATTCGGACTA
ACTGTCATTCGGCCTA
ACTGTCATTCGGACTA
ACTGTCATTCGGACTA
ACTGTCATTCGGACTA
ACTGTCATTCGGACTA
GCTGATAACCGGACTA
ACTGACATTCGGACTA
ACTGTCATTCGGACTA

Consensus= ACTGTCATTCGGACTA

Error Correction

Hamming Distance:
http://en.wikipedia.org/wiki/Hamming_distance

Error Correction

3 different strategies

Error Correction

Kmer-spectra based

Error Correction

Suffix tree based

Error Correction

MSA based

Error Correction

Evaluation of Correction