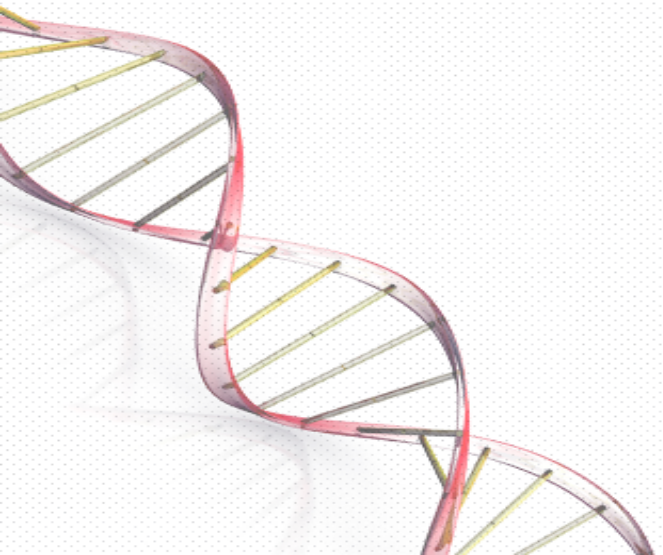
Sequence Database Searching

Hammad Naveed
hammad.naveed@nu.edu.pk

What is database searching?

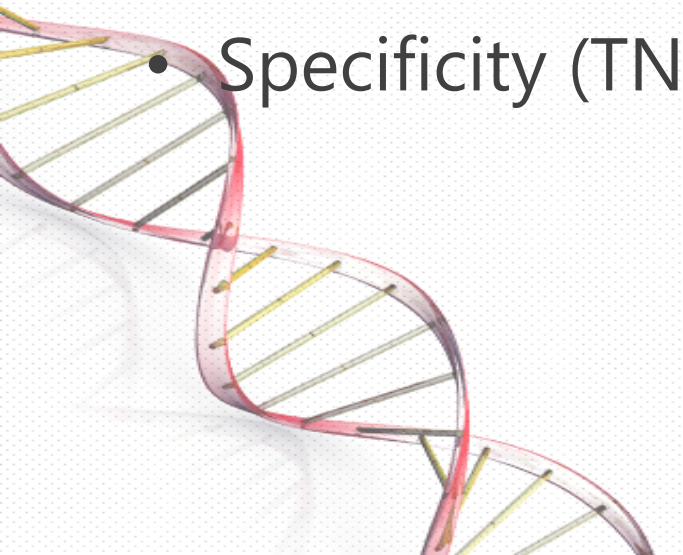
- Goal: find similar (homologous) sequences of a query sequence in a sequence of database

- Input: query sequence & database
- Output: hits (pairwise alignments)



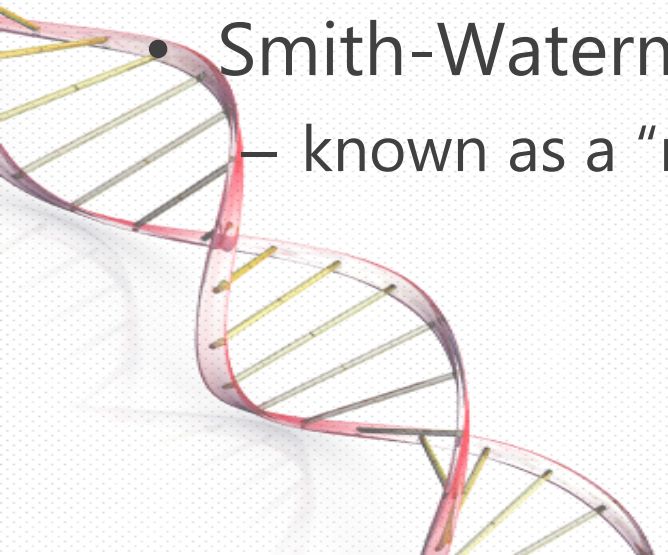
Basics of database searching

- Core: pairwise alignment algorithm
- Speed (fast sequence comparison)
- Relevance of the search results (statistical tests)
- Recovering all information of interest
  - The results depend of the search parameters like gap penalty, scoring matrix.
- Specificity (TN/N) and sensitivity (TP/P)



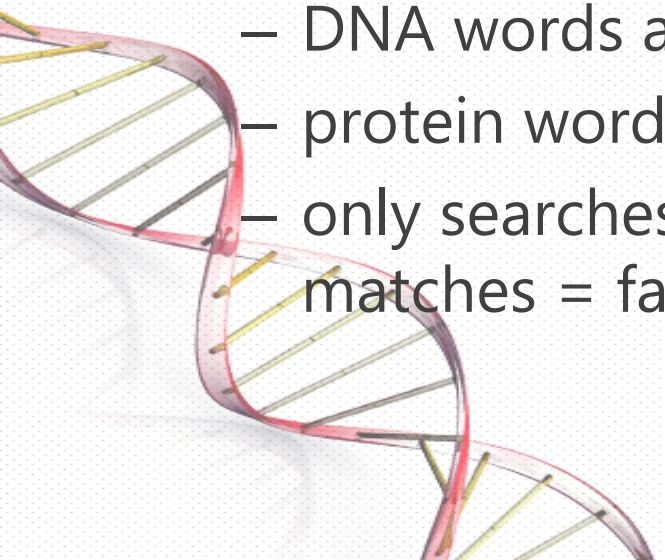
What program to use for searching?

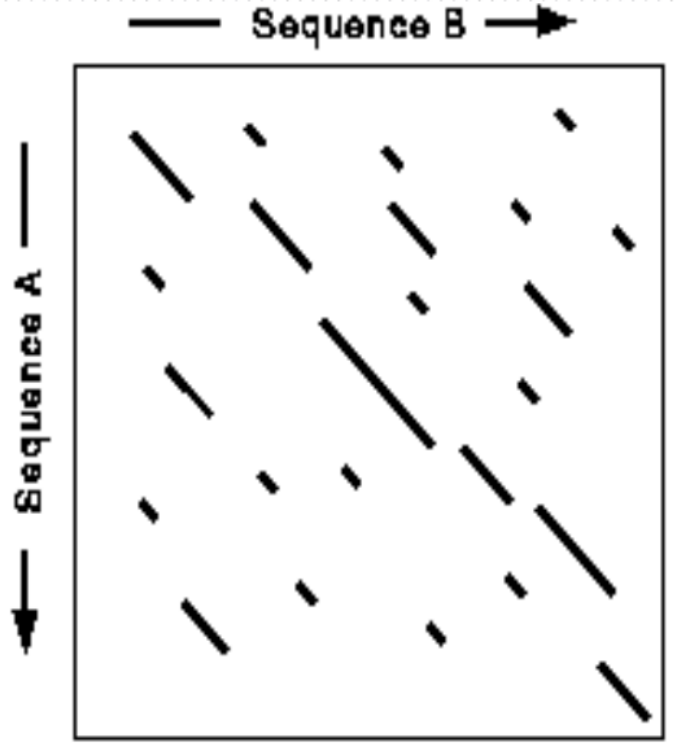
- BLAST
  - fastest and easily accessed on the Web
  - A suite; BLASTP, BLASTN, BLASTX
- FASTA
  - more sensitive for DNA-DNA comparisons
  - FASTX and TFASTX can find similarities in sequences with frameshifts
- Smith-Waterman is slower, but more sensitive
  - known as a “rigorous” or “exhaustive” search



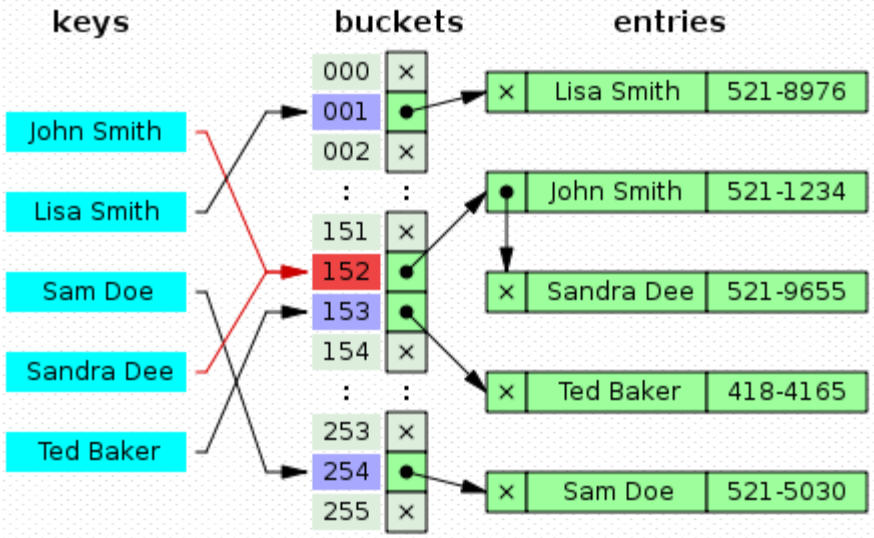
FASTA

- Derived from logic of the dot plot
  - compute best diagonals from all frames of alignment
- Word method looks for exact matches between words in query and test sequence
  - hash tables
  - DNA words are usually 6 bases
  - protein words are 1 or 2 amino acids
  - only searches for diagonals in region of word matches = faster searching



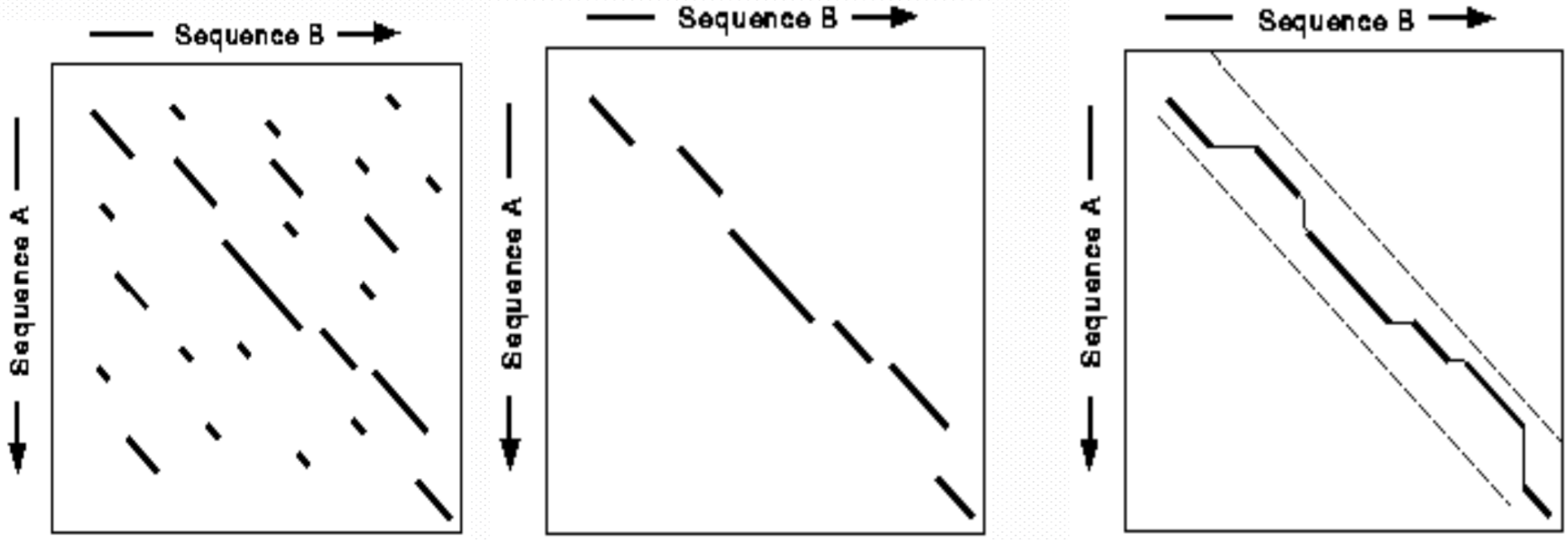


Find runs of identitical words



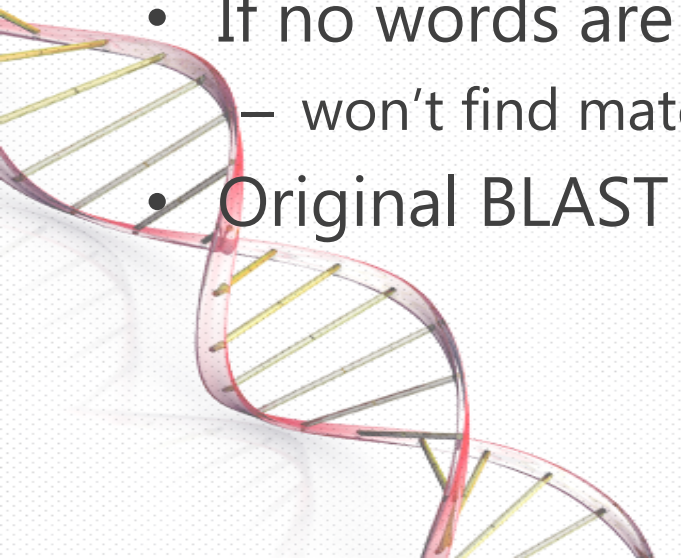
FASTA

- after all diagonals found, tries to join diagonals by adding gaps
- computes alignments in regions of best diagonals
- Dynamic Programming

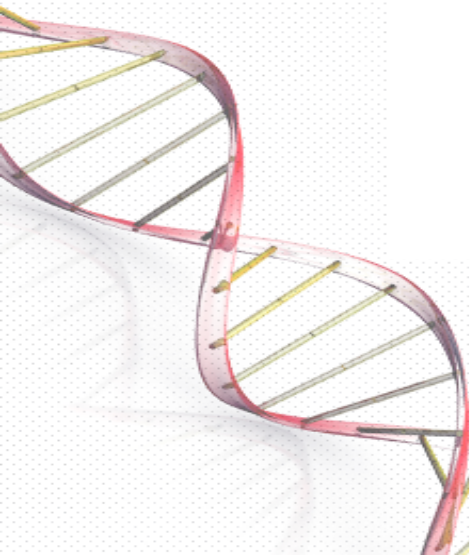


BLAST

- BLAST= Basic Local Alignment Search Tool
- The central idea of the BLAST algorithm is that a statistically significant alignment is likely to contain a high-scoring pair of aligned words
- Uses word matching like FASTA
  - Does not require identical words
  - 3 amino acids or 11 bases
- If no words are similar, then no alignment
  - won't find matches for very short sequences
- Original BLAST does not handle gaps well; the "gapped" blast is better



The BLAST SEARCH ALGORITHM



query word (*W* = 3)

Query: GSVEDTTGSQSLAALLNKCKTPQGQRLVNQWIKQPLMDKNRIEERLNLVEAFVEDAELRQTLQEDL

neighborhood words

PQG	18
PEG	15
PRG	14
PKG	14
PNG	13
PDG	13
PHG	13
PMG	13
PSG	13
PQA	12
PQN	12

neighborhood score threshold (*T* = 13)

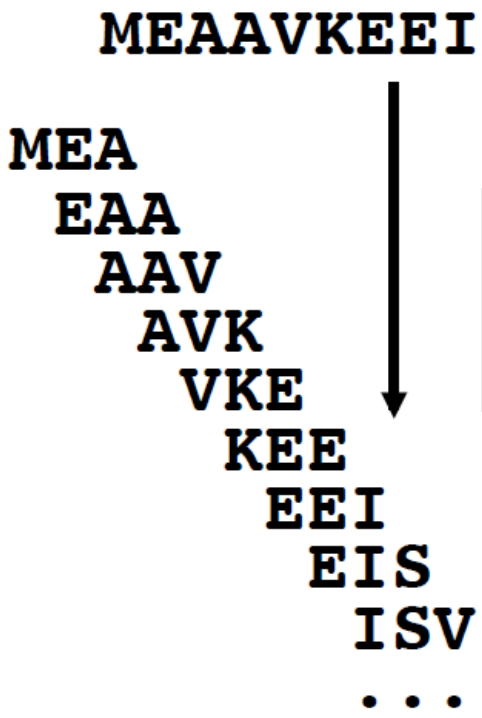
Match Words

Query: 325 SLAALLNKCKTPQGQRLVNQWIKQPLMDKNRIEERLNLVEA 365
+LA++L+ TP G R++ +W+ P+ D + ER + A
Sbjct: 290 TLASVLDCTVTPMGSRMLKRWLHMPVRDTRVLLERQQTIGA 330

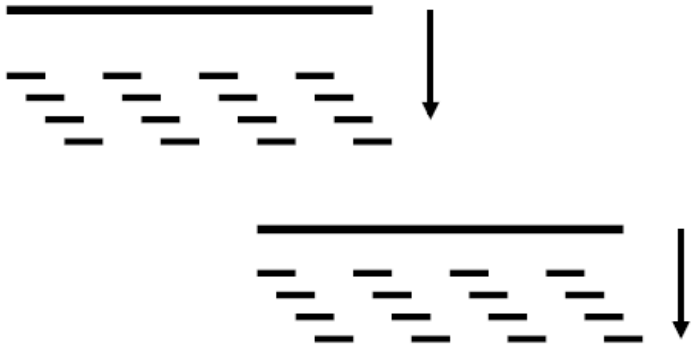
High-scoring Segment Pair (HSP)

Extend Hits

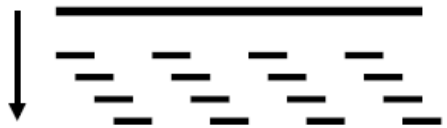
BLAST: Word Matching



Break query into words:



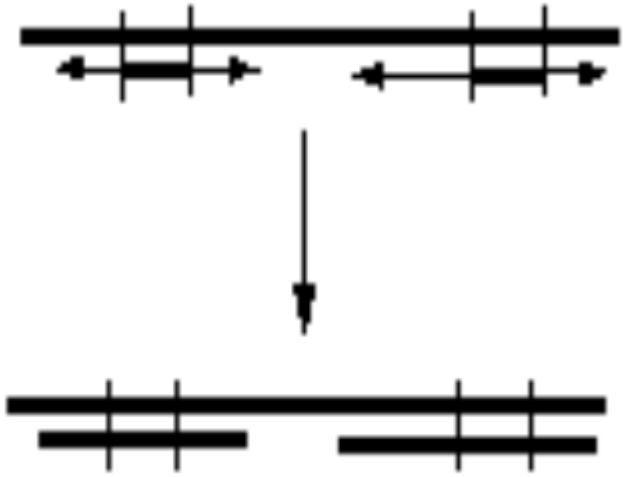
Break database sequences into words:



Compare word lists by Hashing
(allow near matches)

BLAST: Extend Hits

- For each word, extend the alignment in both directions to find alignments that score greater than a threshold of value S



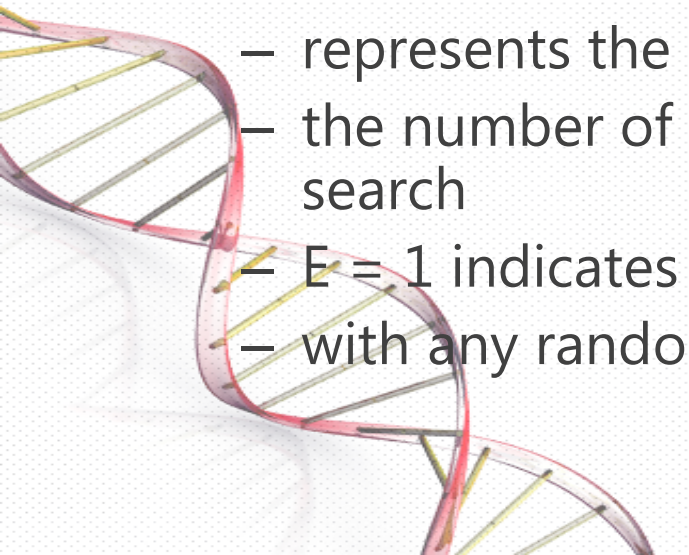
- Use two word matches as anchors to build an alignment between the query and a database sequence

Gapped BLAST algorithm

- The NCBI's BLAST website now uses "gapped BLAST"
- This algorithm is more complex than the original BLAST
- It requires two word matches close to each other on a pair of sequences (i.e. with a gap) before it creates an alignment allow gaps (using Smith-Waterman algorithm)
- Ref: Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. NAR 25(17):3389,1997

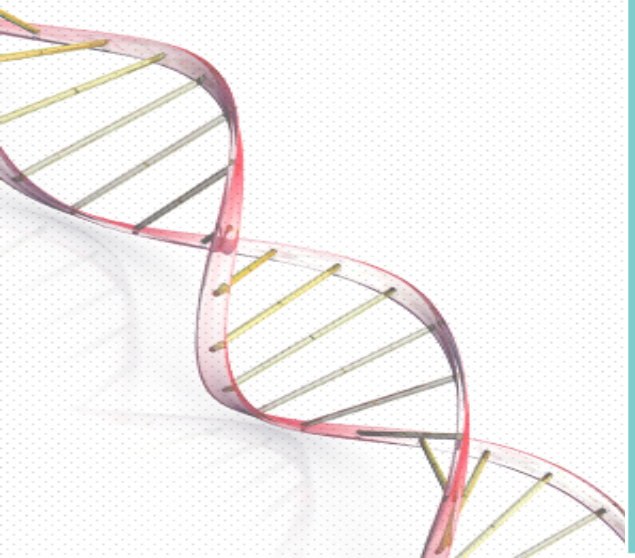
Statistical tests

- Evaluate the probability of an event taking place by chance
- P-value
  - Randomized data
  - Distribution under the same setup
- Example
  - Height of a person, BLAST Score
- E-value
  - represents the likelihood that the observed alignment is due to chance alone.
  - the number of alignments expected by chance (E) during a sequence database search
  - E = 1 indicates that an alignment this good would happen by chance
  - with any random sequence searched against the same database



BLAST Options

Program	Query	Database	Comparison	Common use
blastn	DNA	DNA	DNA level	Seek identical DNA sequences and splicing patterns
blastp	Protein	Protein	Protein level	Find homologous proteins
blastx	DNA	Protein	Protein level	Analyze new DNA to find genes and seek homologous proteins
tblastn	Protein	DNA	Protein level	Search for genes in unannotated DNA
tblastx	DNA	DNA	Protein level	Discover gene structure



BLAST is approximate

- BLAST makes similarity searches very quickly because it takes shortcuts.
  - looks for short, nearly identical “words” (11 bases)

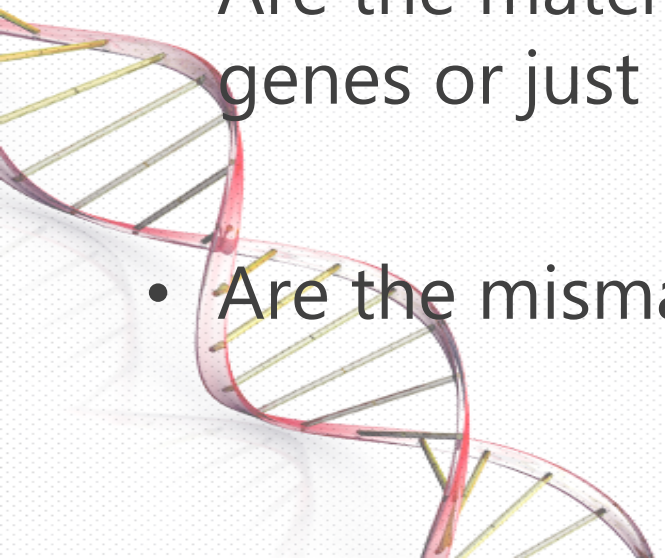
- It also makes errors
  - misses some important similarities
  - makes many incorrect matches
    - easily fooled by repeats or skewed composition

Interpretation of BLAST hits

- very low E values (e-100) are homologs
- moderate E values are related genes
- long list of gradually declining of E values indicates a large gene family
- long regions of moderate similarity are more significant than short regions of high identity

Biological relevance

- Depends on several things….
- Were you looking for a short region of nearly identical sequence or a larger region of general similarity?
- Are the matching regions important structural components of the genes or just introns and flanking regions?
- Are the mismatches conservative ones?



References

- Lecture notes of Colin Dewey @ University of Wisconsin-Madison
- Lecture notes of Arne Elofsson @ Stockholm University
- Lecture notes of Yuzhen Ye @ Indiana University