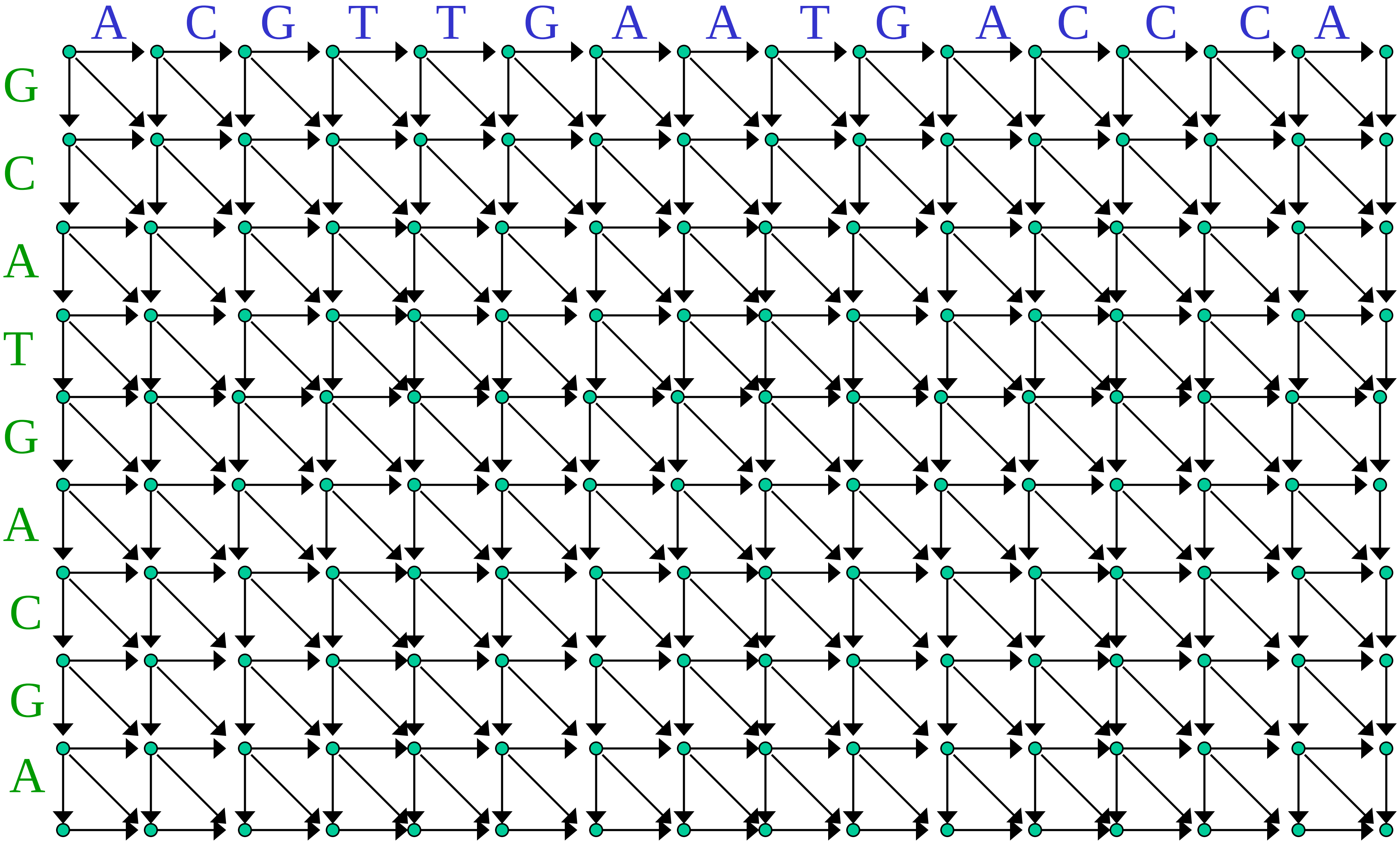
Today’s Lecture

- Smith-Waterman special cases
- Word nucleation algorithms
  - BLAST
- Site models

The *Edit Graph* for a Pair of Sequences

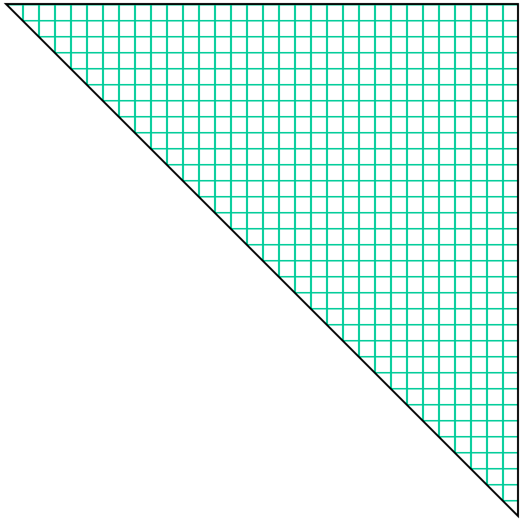


- Find *imperfect internal repeats* by searching edit graph of sequence against itself
  - i.e. the same sequence labels columns and rows

above (& not including) the main diagonal:

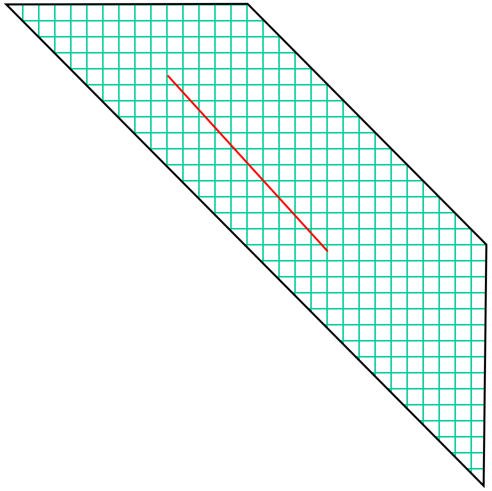
  - if include main diagonal, best path will be identity match to self
  - complexity = $O(N^2)$ where N = sequence length.

Graph for finding imperfect internal repeats:



- Find *short tandem repeats* (e.g. microsatellites, minisatellites):
  - scan a *band* just above main diagonal.
  - Complexity = $O(kN)$ where k is width of the band.
  - Manageable even for large N, if k small.

Graph for finding short tandem repeats:



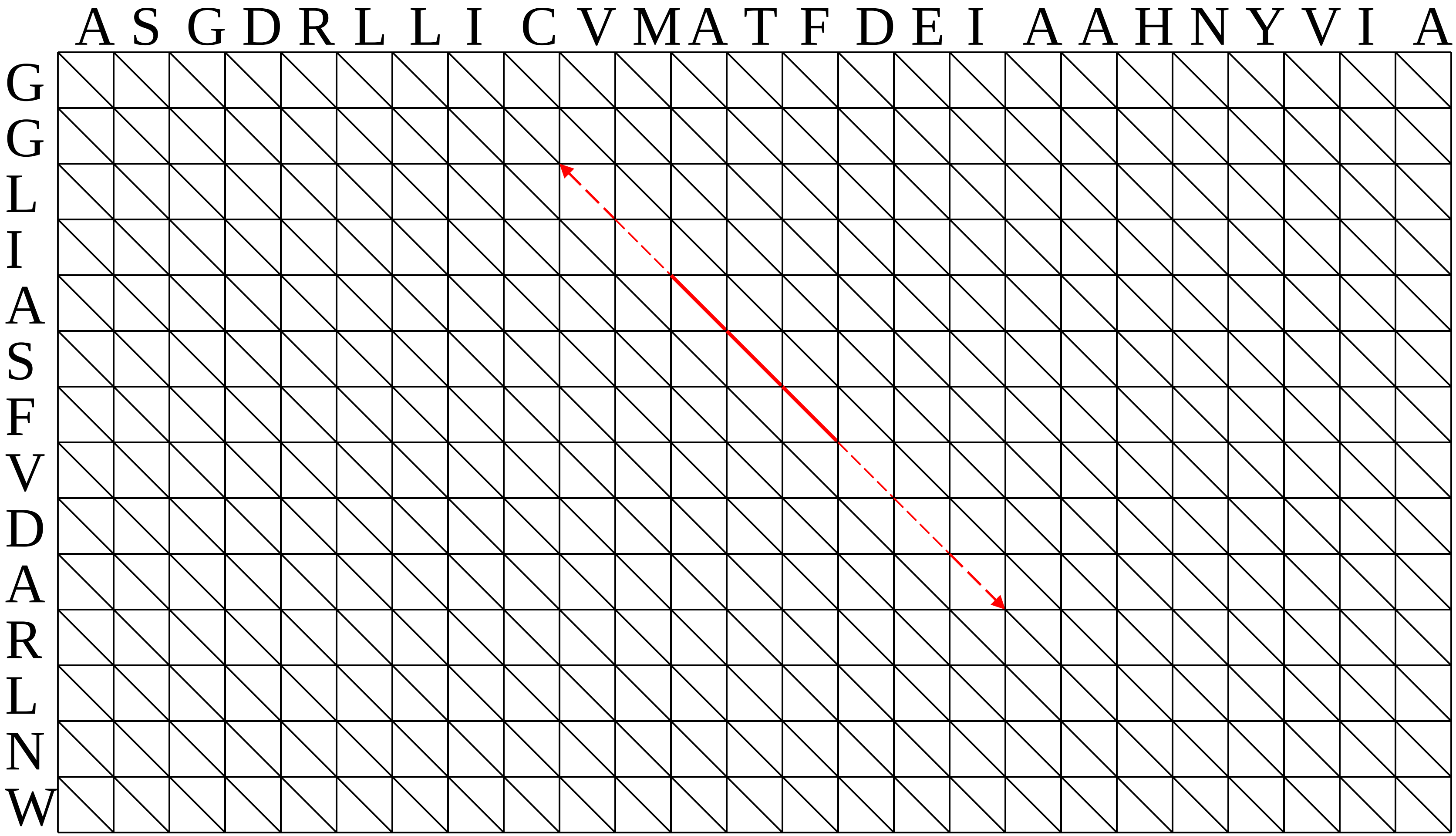
```
ACACACACACACACAC
  ACACACACACACACAC
```

- Other alignment tasks:
  - EST, or cDNA, to genomic sequence (exons)
  - protein to genomic.
- Can solve by variants of Smith-Waterman:
  - e.g. cDNA vs genomic:
    - set moderately large negative penalty for mismatch and for gap opening,
    - 0 for gap extension.
    - issue of proper placement of splice sites ...

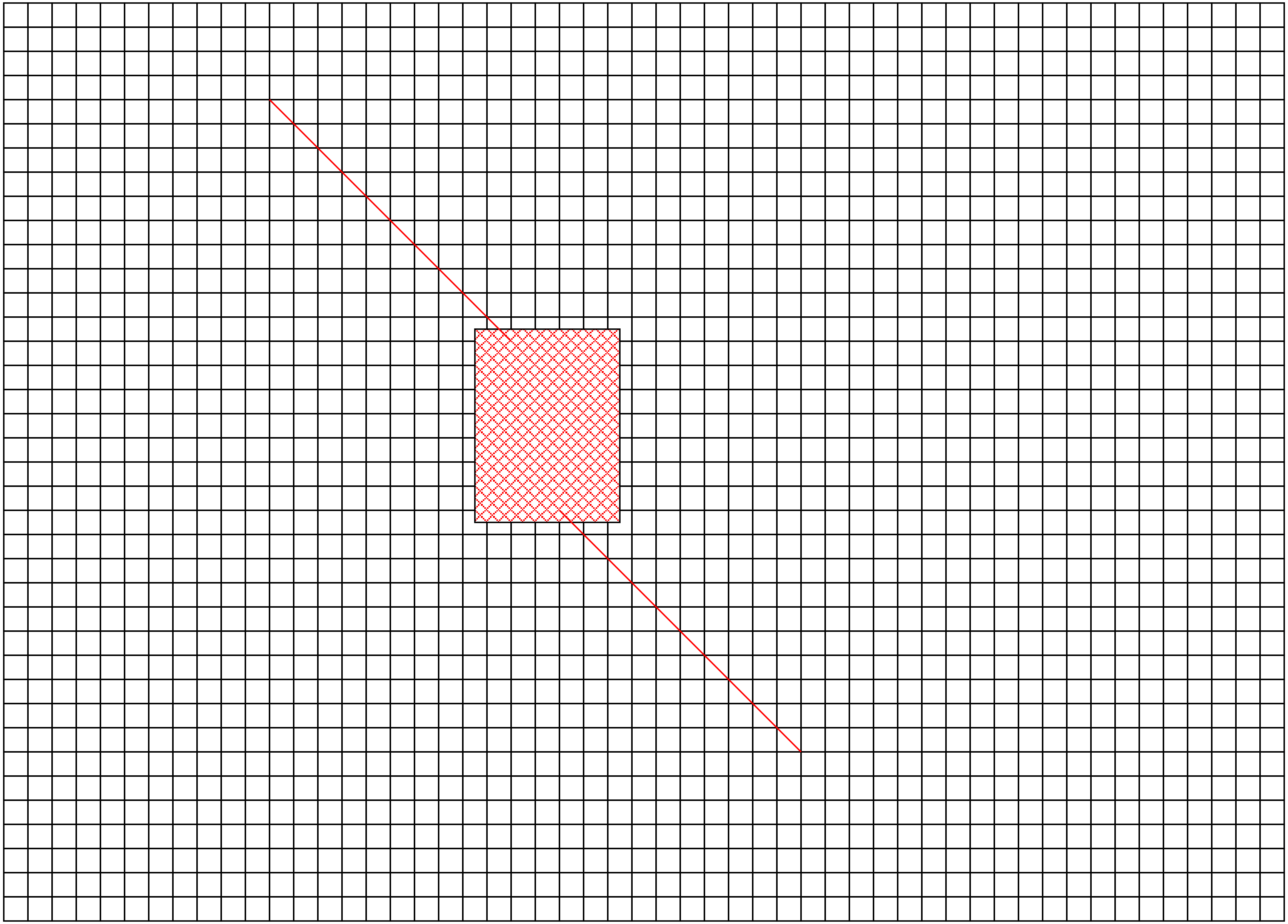
Word Nucleation Algorithms

- Idea: find short (perfect or imperfect) word matches to ‘nucleate’ graph search
  - Each such match defines short *diagonal* path
  - Only search part of graph ‘surrounding’ this path
- BLAST: allow *imperfect* short (e.g. length 3) matches.
  - “*Neighbors*”: set of 3-residue sequences having ≥ min score T against some 3-residue sequence of query
  - Scan database seqs until hit word in neighbor list
  - then do ungapped extension (along diagonal defined by word match)
    - ‘significant’ matches are those with scores ≥ a threshold S
    - Ungapped matches are effective for detecting related proteins:
      - **true protein alignments usually include substantial gap-free regions.**

BLAST: Word Nucleating Alignment


A S G D R L L I C V M A T F D E I A A H N Y V I A
G G L I A S F V D A R L N W

- If find ≥ 2 significant ungapped matches in same seq, expand search to connecting region of matrix, allowing gaps:



Other Word Nucleation Programs

- FASTA:
  - look for clusters of short exact matches, on nearby diagonals;
  - when found, extend to gapped alignment
- *cross_match*:
  - do full search of *bands* around exact matches
- These all still time complexity $O(MN)$
  - because # word matches proportional to MN

but with much smaller constant.

- In database searches, most seqs unrelated to query
- suggests following strategy:
  - Initial rapid pass through database using fast algorithm
    - e.g. just looking for gap-free matches

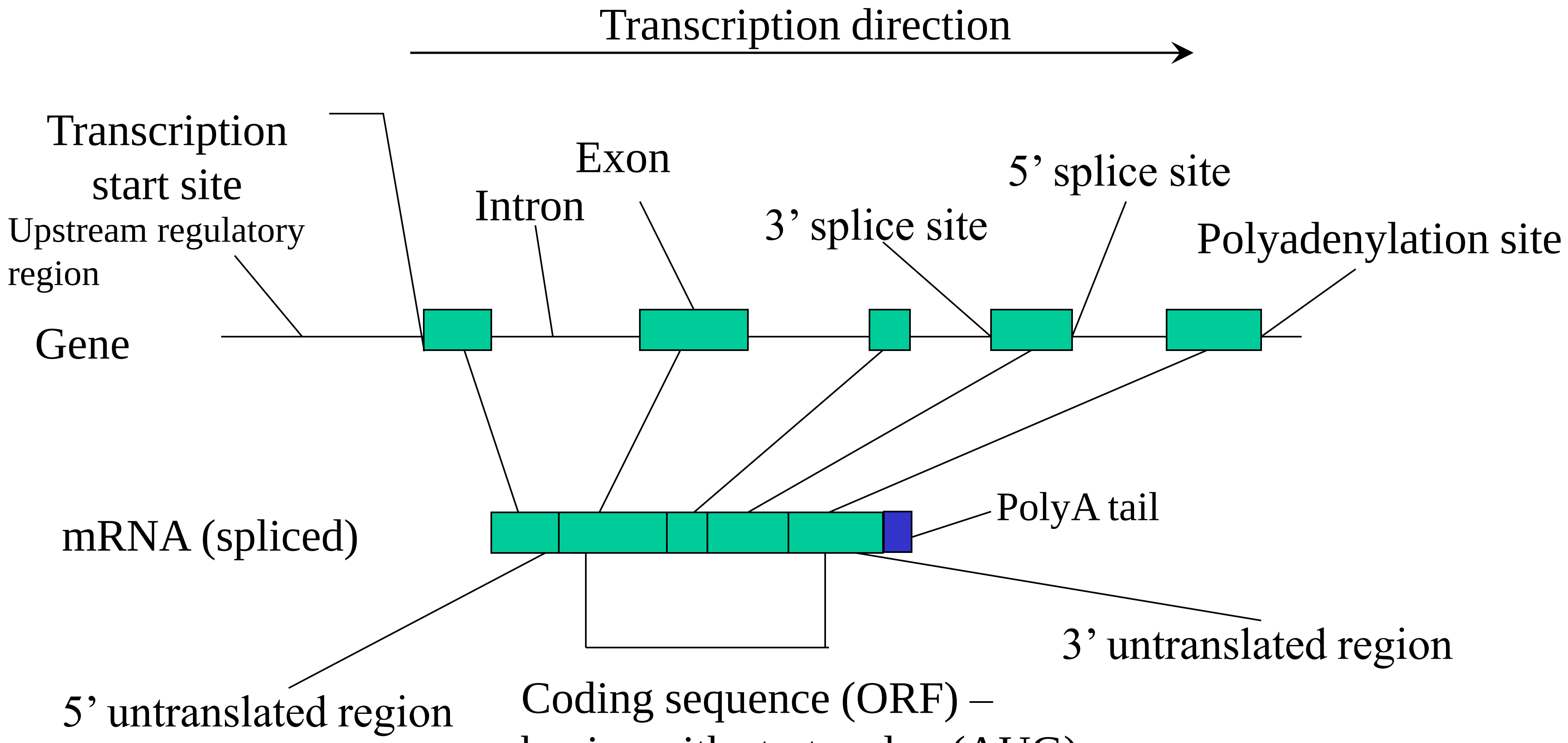
    to get (approximate) score,
  - identify sequences having scores above a threshold
  - use full Smith-Waterman on latter
  - for appropriate (low) threshold can get sensitivity nearly as good as full Smith-Waterman search.

- Important issue: statistical significance for database searches! We will return to this later (Karlin-Altschul theory).

Site Models

- Probability models for short sequences, such as:
  - splice sites
  - translation start sites
  - promoter elements
  - protein “motifs”

(Protein-coding) Gene Structure in Eukaryotes



Coding sequence (ORF) – begins with start codon (AUG), ends with stop codon (UAA, UAG, or UGA)

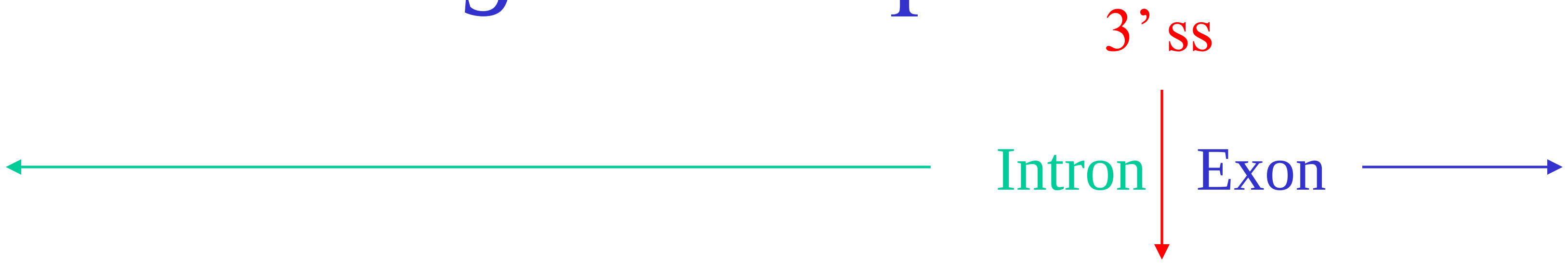
- Assumptions:
  - different examples of site can be aligned *without gaps* (indels) such that tend to have same residues in same positions
  - drop equal freq assumption: allow *position-specific freqs*
  - retain *independence* assumption (for now)

- Applies to short segments (< 30 residues) where
    - precise residue spacing is structurally or functionally important, and
    - certain positions are highly conserved
- Examples:
    - DNA/RNA sequences binding a single protein or RNA molecule
    - Protein internal regions structurally constrained due to folding requirements; or
    - protein surface regions constrained because bind certain ligands

Construction of Site Models

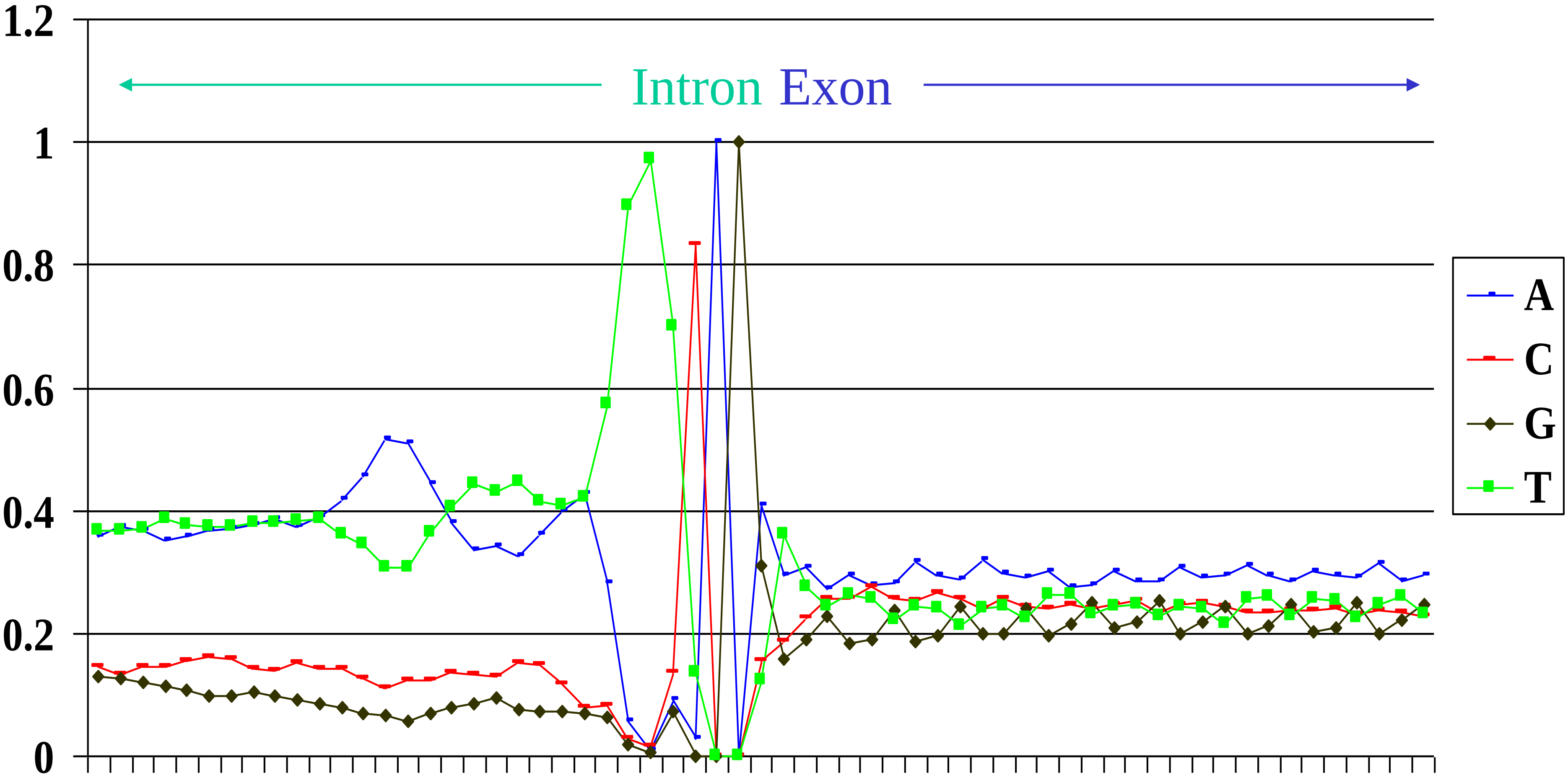
- Collect examples of site
- Align (without gaps)
- Count occurrences of residues at each position
- Convert to frequencies

Nucleotide Counts for 8192 *C. elegans* 3’ Splice Sites

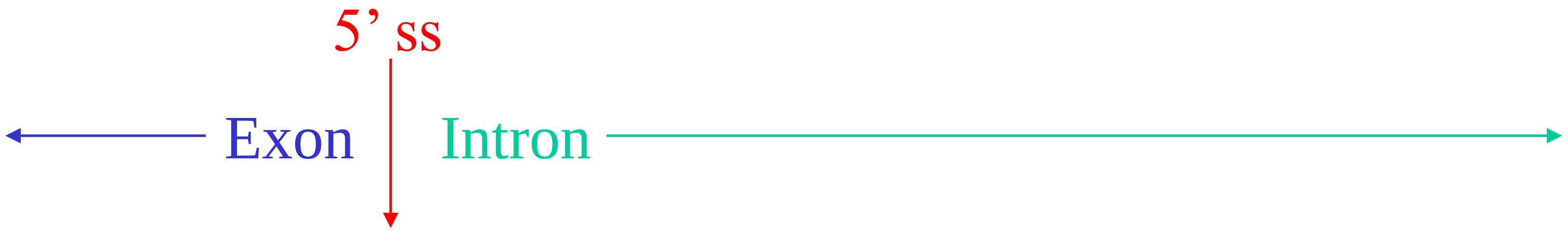


A	3276	3516	2313	476	67	757	240	8192	0	3359	2401	2514
C	970	648	664	236	129	1109	6830	0	0	1277	1533	1847
G	593	575	516	144	39	595	12	0	8192	2539	1301	1567
T	3353	3453	4699	7336	7957	5731	1110	0	0	1017	2957	2264
CONSENSUS	W	W	W	T	T	t	C	A	G	r	w	w
A	0.400	0.429	0.282	0.058	0.008	0.092	0.029	1.000	0.000	0.410	0.293	0.307
C	0.118	0.079	0.081	0.029	0.016	0.135	0.834	0.000	0.000	0.156	0.187	0.225
G	0.072	0.070	0.063	0.018	0.005	0.073	0.001	0.000	1.000	0.310	0.159	0.191
T	0.409	0.422	0.574	0.896	0.971	0.700	0.135	0.000	0.000	0.124	0.361	0.276

3’ Splice Sites – *C. elegans*

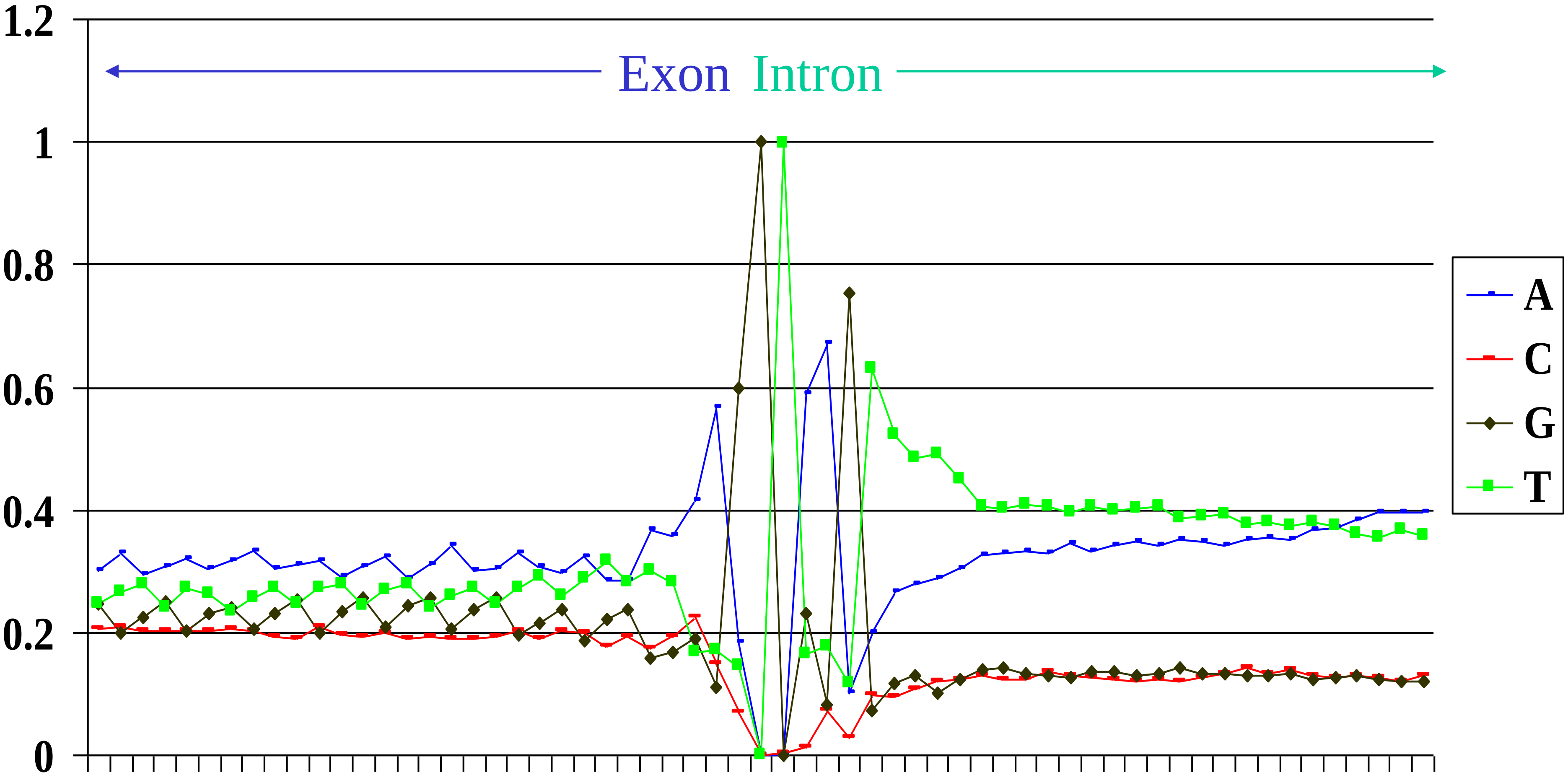


Nucleotide Counts for 8192 *C. elegans* 5’ Splice Sites



A	3404	4644	1518	0	0	4836	5486	837	1632	2189	2278	2355
C	1850	1224	583	0	14	118	588	237	801	771	889	986
G	1562	912	4891	8192	0	1890	672	6164	589	962	1056	827
T	1376	1412	1200	0	8178	1348	1446	954	5170	4270	3969	4024
CONSENSUS	x	a	g	G	T	a	a	g	t	t	w	t
A	0.416	0.567	0.185	0.000	0.000	0.590	0.670	0.102	0.199	0.267	0.278	0.287
C	0.226	0.149	0.071	0.000	0.002	0.014	0.072	0.029	0.098	0.094	0.109	0.120
G	0.191	0.111	0.597	1.000	0.000	0.231	0.082	0.752	0.072	0.117	0.129	0.101
T	0.168	0.172	0.146	0.000	0.998	0.165	0.177	0.116	0.631	0.521	0.484	0.491

5’ Splice Sites – *C. elegans*



Conserved Domain in RecR and Class I Topisomerases

```
RecR   RLAEEKITEVILATNPTVEGEATANYIAELC
RecM   RLQDDQVTEVILATNPNIEGEATAMYISRLL
RecR   RVDDVGITEVIIATDPNTEGEATATYLVRMV
TrsI   IFKENKIDEVIIATDPAREGENIAYKILNQL
TOP1   KQLAEKADHIYLATDLDREGEAIAWRLREVI
ORF1   AELLKQANTIIVATDSDREGENIAWSIIHKA
TOP1   KDALKDADELILATDEDREGKVISWHLLQLL
TOP1   TIFDKRVKTIILATDAAAEGEYIGRNILYRL
TOP3   KREARNADYLMIWTDCDREGEYIGWEIWQEA
TOP3   KRFLHEASEIVHAGDPDREGQLLVDEVLDYL
RGYR   RNLAVEADEVLIGTDPDTEGEKIAWDLYLAL

CONSENSUS  xxxxxxxxxU&uatDxxxEGexxxxxUxxxu
```

Consensus key:

Uppercase: all residues chemically similar

lowercase: most are

U,u: bulky aliphatic (I,L,V)

&: bulky hydrophobic (I,L,V,M,F,Y,W)

From RL Tatusov, SF Altschul, and EV Koonin, PNAS 91: 12091-12095

Probability Models for Sites (assuming independence!)

- For each position *i*, $1 \leq i \leq n$, let P_i be a prob dist'n on the alphabet of residues
  - e.g. constructed using counts at that position in a sample of sites.
  - $P_i(r)$ for each residue *r* is the probability that *r* occurs at position *i* in a sequence.
- Prob dist'n *P* on the space *S* of sequences of length *n* is defined by

$$P(s) = \prod_{1 \leq i \leq n} P_i(s_i)$$

where $s = s_1 s_2 \ldots s_n$

Zero Probabilities

- If $P_i(r)$ = 0 for some *i* and *r*, then $P(s)$ = 0 for some sequences.
  - may or may not be desirable
- If due to failure to observe residue because of small sample size,
  - should perform “small-sample correction” to change $P_i(r)$ to a small non-zero value.
  - usually done by adding ‘pseudocounts’ to each value in the counts matrix;
    - e.g. add 1 to each cell (has justification in Bayesian statistics)
  - Particularly an issue with proteins, due to larger alphabet size.
- If reflects real biological constraints
  - then leave as 0.
  - e.g. requirement for G at position +1 (first intronic base) in 5’ss