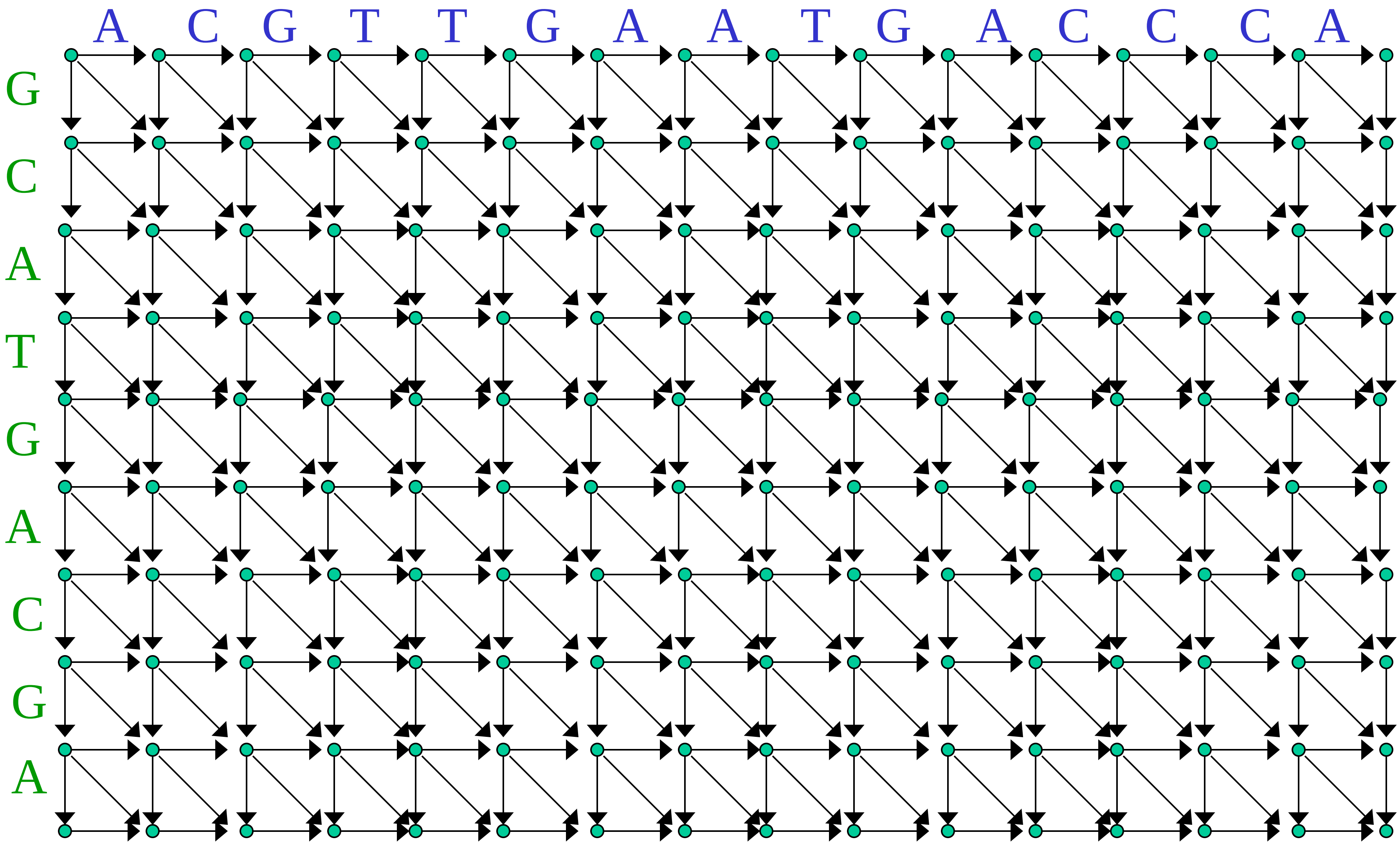
Today’s Lecture

- Smith-Waterman special cases
- Word nucleation algorithms
  - BLAST
- Probability models for sequences

The *Edit Graph* for a Pair of Sequences

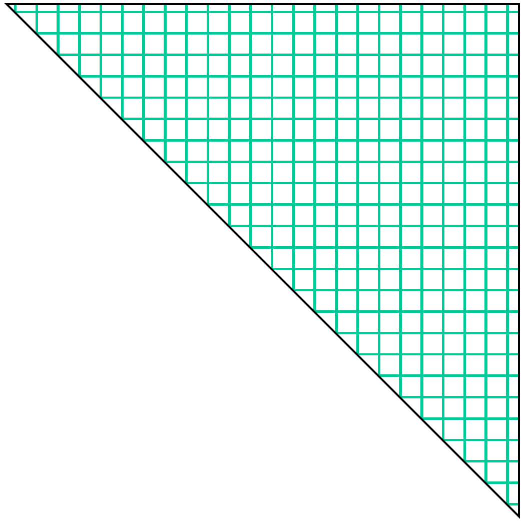


- Find *imperfect internal repeats* by searching edit graph of sequence against itself
  - i.e. the same sequence labels columns and rows

above (& not including) the main diagonal:

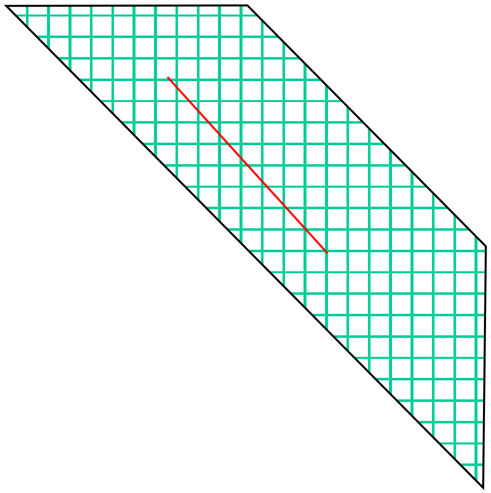
  - if include main diagonal, best path will be identity match to self
  - complexity = $O(N^2)$ where N = sequence length.

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:



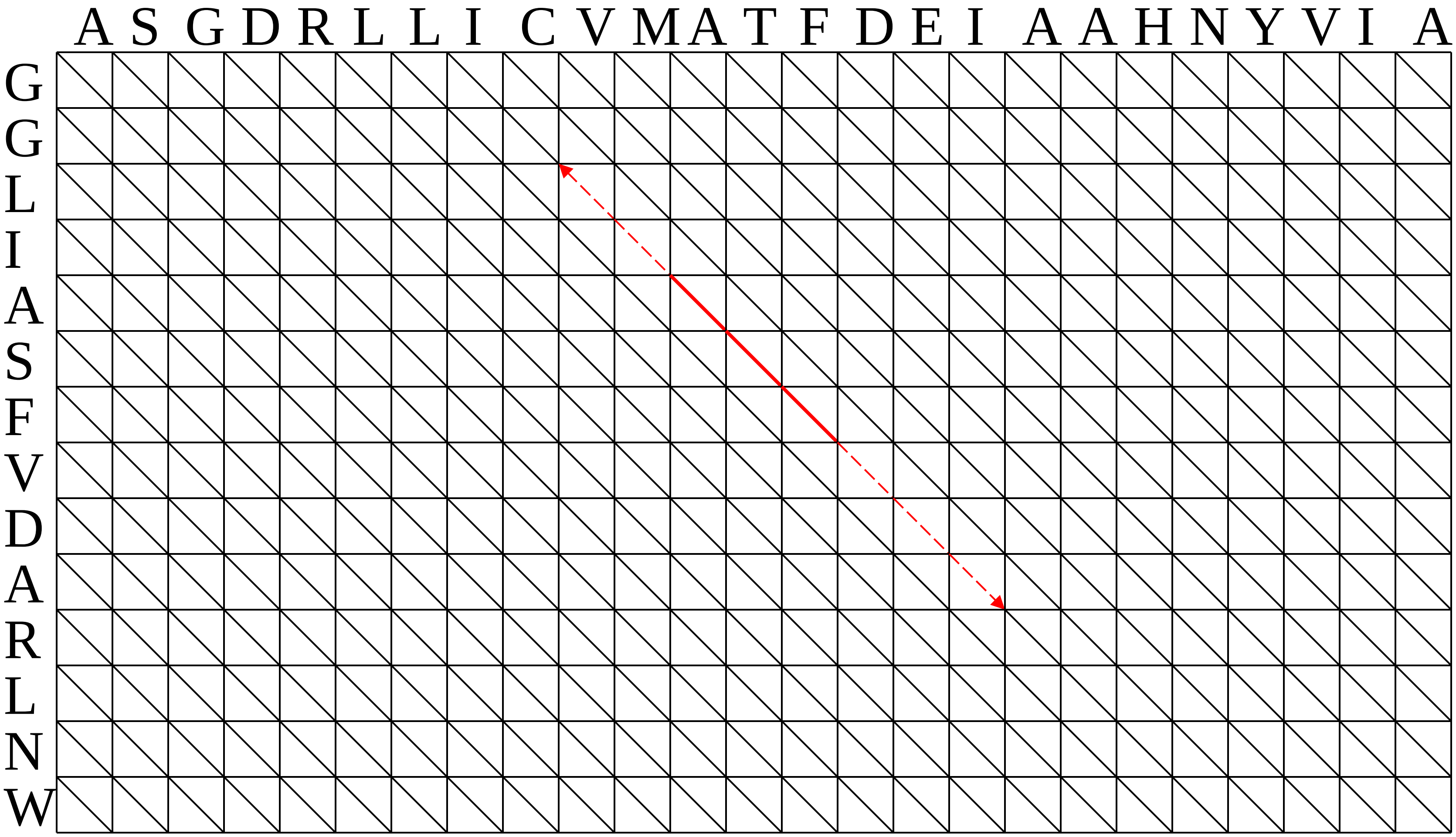
```
ACACACACACACACAC
  ACACACACACACACAC
```

- 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 ...

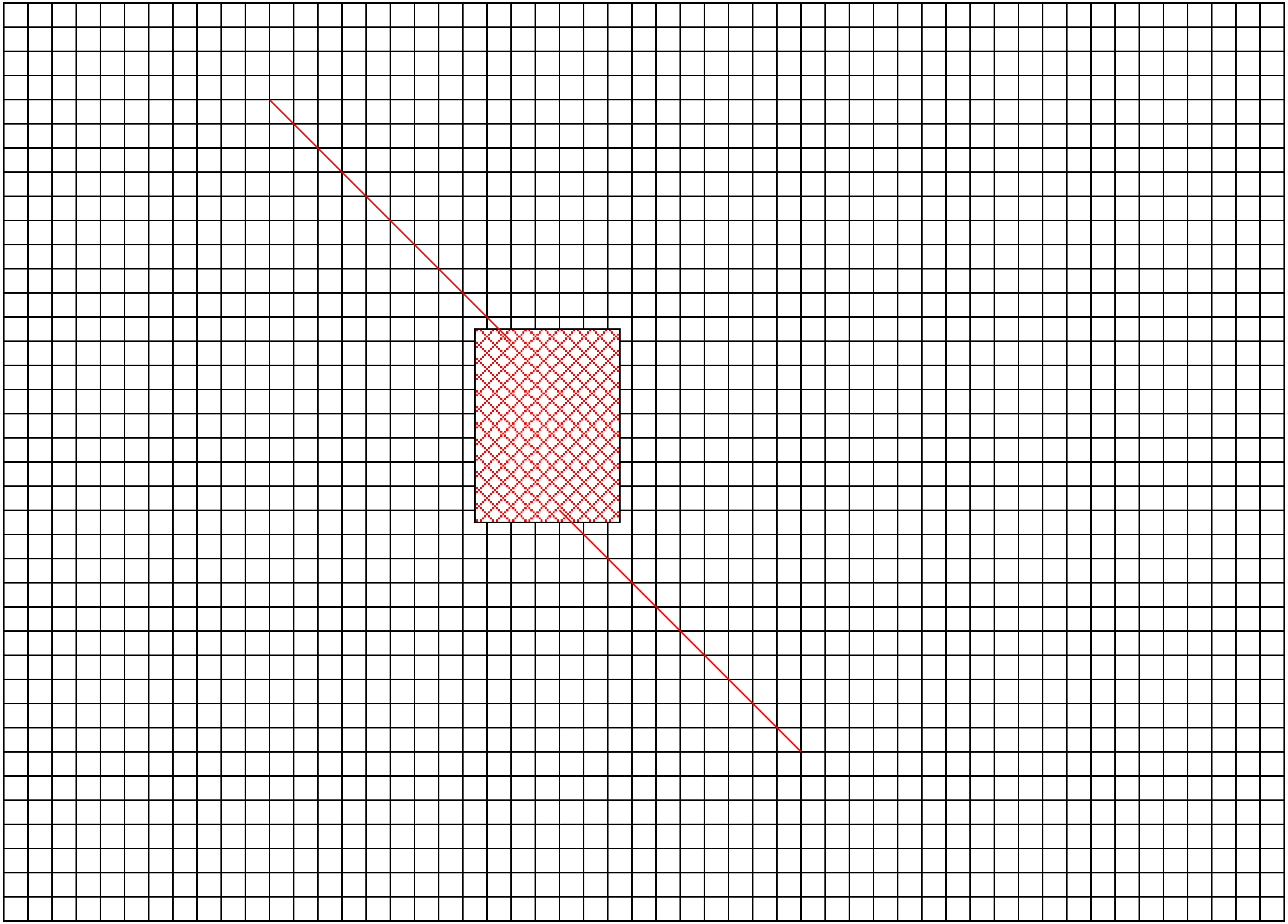
Word Nucleation Algorithms

- 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.

Biology involves *probabilities*, at several levels:

- Fundamental laws of nature
- Mutations (imperfect replication)
- Transmission of DNA from parent to offspring in populations of individuals
- Random aspects of environment

Key Physical Laws Governing Living Organisms

- Individual atoms & molecules:
  - quantum mechanics / quantum electrodynamics
- Systems of molecules:
  - statistical mechanics / 2d law of thermodynamics

These fundamental laws are essentially probabilistic!

“*The true logic of this world is in the calculus of probabilities*” – James Clerk Maxwell

“*I cannot believe that God plays dice with the cosmos*” – Albert Einstein; nonetheless two of his three great 1905 papers dealt with statistical aspects of nature (photoelectric effect & Brownian motion)!

Probability Models of Sequences

- Sample questions in genome sequence analysis:
  - Is this sequence a splice site?
  - Is this sequence part of the coding region of a gene?
  - Are these two sequences evolutionarily related?
  - Does this sequence show evidence of selection?
- Computational analysis can’t answer:
  - only generates *hypotheses*

    which must ultimately be tested by experiment.
- *But* hypotheses should
  - have some reasonable chance of being correct, and
  - carry indication of reliability.

- We use *probability models* of sequences to address such questions.
- Not the only approach, but usually the most powerful, because
  - seqs are products of evolutionary process which is *itself* probabilistic
  - want to detect biological “signal” against “noise” of background sequence or mutations.

- “*All models are wrong; some models are useful.*” – George Box
- “*What is simple is always wrong. What is not is unusable.*” – Paul Valery

Basic Probability Theory Concepts

- A *sample space* S is set of all possible outcomes of a conceptual, repeatable experiment.
  - $|S| < \infty$ in most of our examples.
  - e.g. S = all possible sequences of a given length.
- Elements of S are called *sample points*.
  - e.g. a particular seq = outcome of "experiment" of extracting seq of specified type from a genome.
- A *probability distribution* P on S assigns non-neg real number $P(s)$ to each $s \in S$, such that

$$\Sigma_{s \in S} P(s) = 1$$

(So $0 \leq P(s) \leq 1 \;\; \forall s$)

  - Intuitively, $P(s)$ = fraction of times one would get s as result of the expt, if repeated many times.

- A *probability space* (S,P) is a sample space S with a prob dist'n P on S.
- Prob dist'n on S is sometimes called a *probability model* for S, particularly if several dist'ns are being considered.
  - Write models as M_1, M_2 , probabilities as $P(s \mid M_1)$, $P(s \mid M_2)$.
  - e.g.
    - M_1 = prob dist'n for splice site seqs,
    - M_2 = prob dist'n for "background" (arbitrary genomic) seqs.

Basic Probability Theory Concepts (cont'd)

- An *event* E is a criterion that is true or false for each $s \in S$.
  - defines a subset of S (sometimes also denoted E).
  - $P(E)$ is defined to be $\Sigma_{s|E \text{ is true}} P(s)$.
- Events E_1, E_2, … , E_n are *mutually exclusive* if no two of them are true for the same point;
  - then $P(E_1 \text{ or } E_2 \text{ or } \ldots \text{ or } E_n) = \Sigma_{1 \le i \le n} P(E_i)$.
- If E_1, E_2, … , E_n are also *exhaustive*, i.e. every s in S satisfies E_i for some i, then $\Sigma_{1 \le i \le n} P(E_i) = 1$.

- For events E and H, the *conditional probability* of E given H, is

$$P(E \mid H) \equiv P(E \text{ and } H) \,/\, P(H)$$

(= prob that both E and H are true, given H is true)

  – undefined if $P(H) = 0$.

- E and H are (*statistically*) *independent* if

$$P(E) = P(E \mid H)$$

(i.e. prob. E is true doesn’t depend on whether H is true);

or equivalently

$$P(E \text{ and } H) = P(E)P(H).$$

Probabilities on Sequences

- Let S = space of DNA or protein sequences of length n. Possible assumptions for assigning probabilities to S:
  - *Equal frequency assumption*: All residues are equally probable at any position;
    - $P(E_r^{(i)}) = P(E_q^{(i)})$ for any two residues r and q,
      - where $E_r^{(i)}$ means residue r occurs at position i, then
    - Since for fixed i the $E_r^{(i)}$ are mutually exclusive and exhaustive,

$$P(E_r^{(i)}) = 1 / |A|$$

where A = residue alphabet

$$P(E_r^{(i)}) = 1/20 \text{ for proteins, } 1/4 \text{ for DNA).}$$

  - *Independence assumption*: whether or not a residue occurs at a given position is independent of residues at other positions.

- Given above assumptions, the probability of the sequence

  s = ACGCG

(in the space *S* of all length 5 sequences) is calculated by considering 5 events:

- Event 1 is that first nuc is A. Probability = .25.
- Event 2 is that 2d nuc is C. Probability = .25.
- Event 3 is that 3d nuc is G. Probability = .25.
- Event 4 is that 4th nuc is C. Probability = .25.
- Event 5 is that 5th nuc is G. Probability = .25.

By independence assumption, prob of all 5 events occurring is the product $(.25)^5 = 1/1024$.

Since *s* is the only sequence satisfying all 5 conditions, *P*(*s*) = 1/1024.

- More generally, under equal freq and indep assumptions,

  prob of nuc sequence of length n = $.25^n$,

  prob of protein sequence of length n = $.05^n$

  in the space S of length n sequences.