Today's Lecture

- PhastCons
- Karlin-Altschul theory

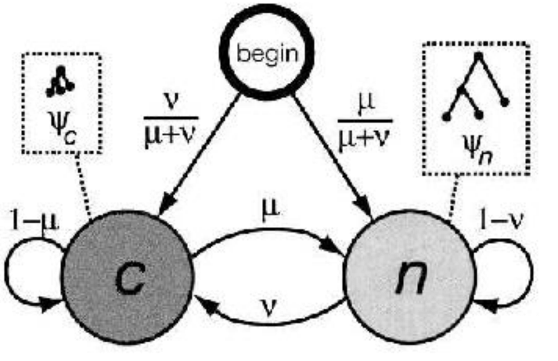
Notation

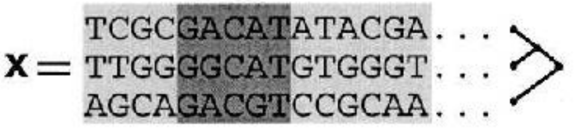
- $\mu = a_{cn}$, $\omega = 1/\mu$ (expected length of conserved elt)
- $\nu = a_{nc}$
- expected 'coverage' γ (frac of genome that is conserved):

  = Elen (cons seg) / (Elen(cons seg) + (Elen(neut seg))

  $= \; (1/\mu) / (1/\mu + 1/\nu)$

  $= \; \nu / (\mu + \nu)$





- L_{min}: expected min length of a conserved segment that could appear in a Viterbi path
- at L_{min} ,

  expected loglike of staying in state n

  = expected loglike of switching to c & back again, so

$$(L_{\min}+1)\log(1-\nu)+L_{\min}\sum_x P(x|\psi_c)\log P(x|\psi_n)$$

$$= \log\nu + \log\mu + (L_{\min}-1)\log(1-\mu) + L_{\min}\sum_x P(x|\psi_c)\log P(x|\psi_c)$$

- $L_{\min} = \dfrac{\log \nu + \log \mu - \log(1-\nu) - \log(1-\mu)}{\log(1-\nu) - \log(1-\mu) - H(\psi_c \| \psi_n)}$

- where

$$H(\psi_c||\psi_n) = \sum_x P(x|\psi_c) \log \frac{P(x|\psi_c)}{P(x|\psi_n)}$$

= rel entropy of *c*-state emission prob dist'n w.r.t.

n-state dist'n

- PIT (phylogenetic information threshold)

= $L_{\min} H(\psi_c||\psi_n)$

= 'expected min amt of phylogenetic info required to predict conserved element'

- Final param estimates (for vertebrates):
  - $\gamma = 0.265$
  - $\omega = 12.0$ bp
  - $H(\psi_c \| \psi_n) = .608$ bits / site
  - $L_{min} = 16.1$ bp
  - PIT $= L_{min} H(\psi_c \| \psi_n) = 9.8$ bits

Group	Method	Total no.[a]	Ave. len.[b]	Cov.[c]	CDS cov.[d]	μ	ν	ω	γ	L_{min}
vert.	MLE	561,103	216.1	4.2%	68.8%	0.018	0.004	55.4	0.191	30.4
	55%	1,058,855	75.3	2.8%	56.8%	0.125	0.029	8.0	0.187	12.9
	65%[e]	1,157,180	103.5	4.2%	66.1%	0.083	0.030	12.0	0.265	16.0
	75%	1,381,978	167.5	8.1%	76.6%	0.043	0.031	23.0	0.415	22.6

Group	Method	Total no.[a]	Ave. len.[b]	Cov.[c]	CDS cov.[d]	CDS frac.[e]	$H(\psi_c\|\|\psi_n)$	L_{min}
vert.	65%	1,157,180	103.5	4.2%	66.1%	18.0%	0.611	16.0
	4d	797,777	109.3	3.0%	64.2%	24.0%	0.854	11.0

Estimating false positive rates

- simulate 1 Mb alignment
  - by sampling 4D sites (with replacement) from aligned CDSs
  - caveat: these not typical of all neutral sites!
- predict cons elts (using prev param estimates)
- frac of bases in cons elts:

Group	65%	75%	MLE
vertebrate	0.00279[a]	0.00362	0.00005
insect	0.00286	0.01026	0.00152
worm	0.00000	0.00000	0.00000
yeast	0.00006	0.00042	0.00023

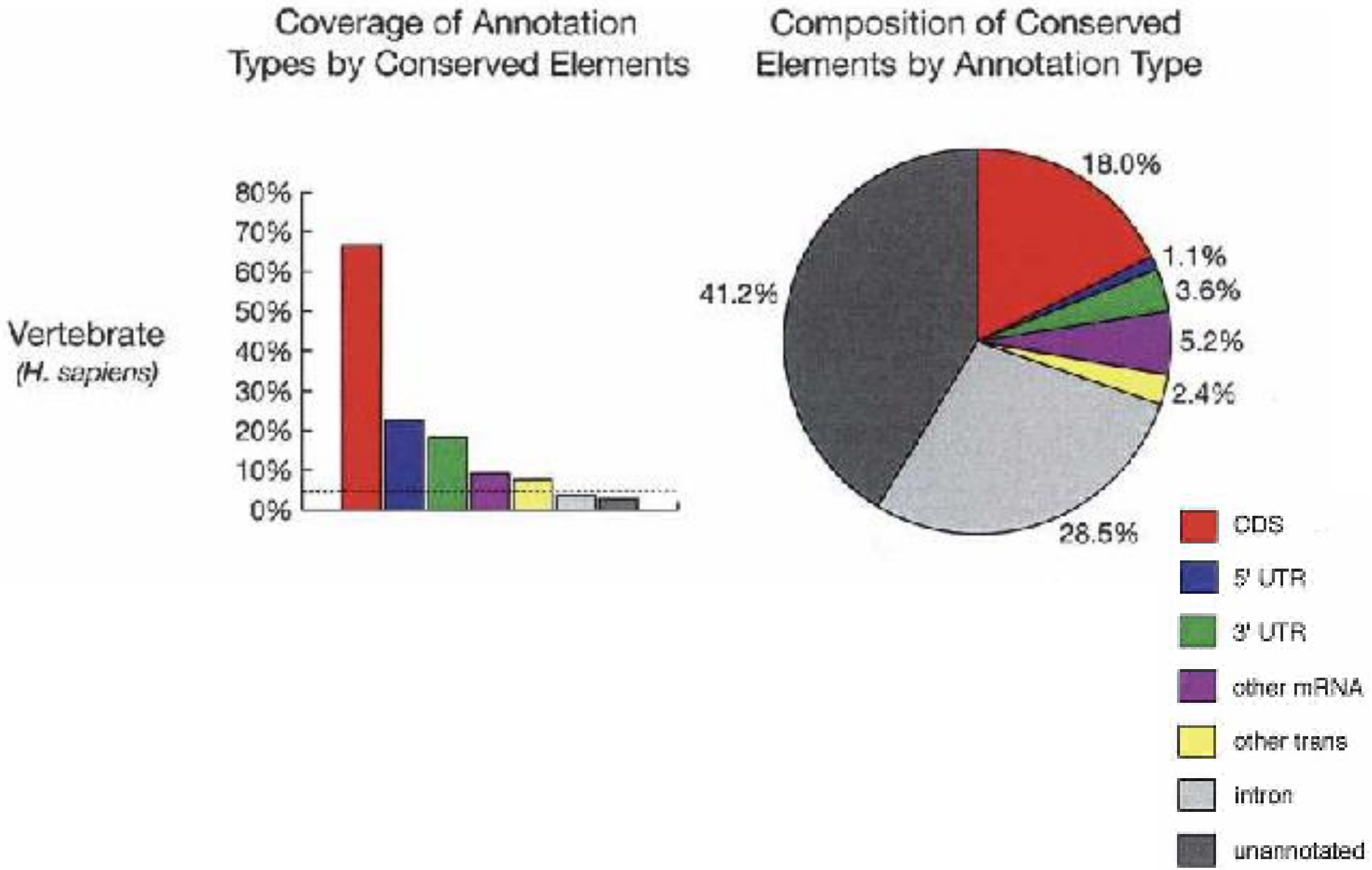
- does not address (important) issue of rate of false positive bases within, or flanking, true conserved elements
- also: genes more G+C rich than genome average, & have somewhat higher mutation rate (due in part to more frequent CpGs)

  $\Rightarrow$ *underestimating* false pos rate
- also: randomization procedure destroys underlying mutation rate variation

  $\Rightarrow$ *underestimating* false pos rate

Characteristics of phastCons predicted conserved elements

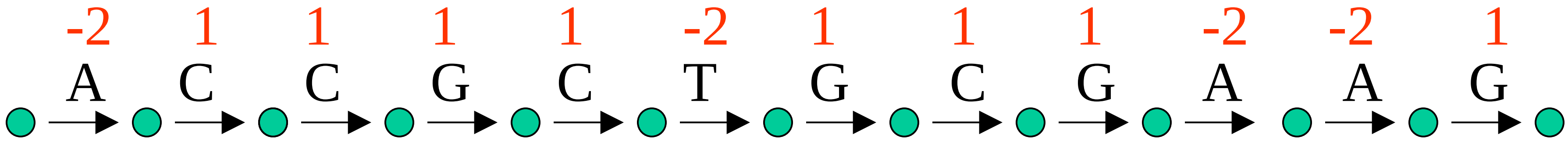
- 1.18 million elements
- constitute 4.3% of human sequence
  - 66% of coding bases
    - 88% of coding exons overlap predicted elt
  - 23% of 5'UTR bases
    - 63% of exons
  - 18% of 3'UTR bases
    - 64% of exons
  - 42% of RNA gene bases
    - 56% of genes
  - 3.6% of intronic bases
  - 2.7% of intergenic bases
  - < 1% of mammalian 'ancestral repeats' (ARs)



from Siepel A. *et al.* (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* 15:1034-50.

Context for Karlin-Altschul Theory for Maximal Segment Analysis

- Linked list, with labels attached to edges, e.g.
  - a sequence graph: labels = sequence residues
  - (ungapped) aligned pair of seqs: labels = possible alignment columns (pairs of residues)
- edge weights depend only on labels:
  - each label is assigned a weight $W(s) = w_s$



- in backgd model, each label s occurs with probability $P(s) = p_s$ where
  - P = prob dist'n on sample space S = {labels}

Methods for Computing Statistical Significance of Maximal Segment Scores

1. exact prob dist'n
2. approximate formula (Karlin-Altschul)
3. from simulated sequences
4. from real biological 'background' sequences
   – i.e. not having feature in question

1, 2, 3 require prob model approximating biological reality; 4 requires an appropriate dataset

2 is faster than 1 or 3, but involves add'l approximations (ignores 'edge effects')

1 requires more complex algorithm

Exact Score Dist'n for Segments in WLLs

- Exact score dist'n (following proof allows position-specific scores and probabilities):
  - Let $P_{k,m}^{(i)}$ = prob that :
    - highest-scoring path *ending at position i* has score *k*, *and also*
    - highest scoring path *ending at any pos'n* $\leq i$ has score *m*
  - special cases:
    - $P_{k,m}^{(i)} = 0$ if $k < 0$ or $m < k$;
    - $P_{0,0}^{(0)} = 1$,
    - $P_{k,m}^{(0)} = 0$ if k or $m \neq 0$
  - dist'n of maximum score is $P_m = \sum_{k \leq m} P_{k,m}^{(N)}$.
    (N = seq length)

- Algorithm to compute $\{P_{k,m}^{(i)}\}$ from $\{P_{k,m}^{(i\text{-}1)}\}$:
  - If $0 < k < m$
    - ($\Rightarrow$ best path ending at position i cannot start at i, and best path ending at position $\leq i - 1$ must have score $= m$)

    then $P_{k,m}^{(i)} = \sum_j p_j^{(i)} P_{k-j,m}^{(i-1)}$
  - if $0 < k = m$
    - ($\Rightarrow$ best path ending at position $\leq i - 1$ may have score $\leq m$)

    then $P_{k,m}^{(i)} = \sum_j p_j^{(i)} \sum_{n \leq m} P_{k-j,n}^{(i-1)}$
  - $P_{0,m}^{(i)} = \sum_j p_j^{(i)} \sum_{n \leq -j} P_{n,m}^{(i-1)}$
  - stop when i reaches N

- Can incorporate Markov chain dependencies in sequence probs:
  - just keep track of preceding residue r as well as k,m : $P_{r,k,m}^{(i)}$.
- Reduce required memory by truncating for large m, with appropriate modifications.
- Would like to have generalization to arbitrary DAG (e.g. edit graphs for sequence alignment)!
  - Difficult, because $P_{k,m}^{(v)}$ not independent for different parent vertices v

Why Is *Approximation* to Exact Score Distribution of Interest?

- faster to compute: useful for database searches
- gives better intuition for score behavior
- *Form* of approximation extends to other situations
  - e.g. gapped alignments

where exact dist'n currently unavailable

Approximate Score Distribution for High-Scoring Segments in WLLs: Karlin-Altschul theory

- Main reason why BLAST is most widely used computational biology tool!
- Ideas closely related to
  - classical random walk and gambler's ruin problems in probability theory
    - (cf. W. Feller, *An Introduction to Probability Theory and Its Applications*),
  - sequential sampling in statistics

Karlin-Altschul Theory

- ***Scoring systems*: What is appropriate scoring system (choice of edge weights) for detecting 'target' features in a biological sequence?
  - Answer: if symbol *r* occurs with freq
    - t_r in target segments, and
    - b_r elsewhere ('background')

  the best score is

$$s_r = \log(t_r / b_r)$$

    - N.B. requires knowing (approximately) these frequencies!
  - Moreover, any 'interesting' scoring system can be expressed in above form

- ***Statistical Significance*:**

  Expected # maximal segs of score ≥ S in 'backgd' sequence is

$$NKe^{-\lambda S}$$

  where

  - λ is a scaling factor to convert scores to LLR scale,
  - N = sequence length
  - K is constant (depends on scoring system, but not on S or N)

- (Is above also true for maximal D-segments?)

Scoring systems (Choice of edge weights in WLLs):

- assume *position independent* scores w, probabilities p_w
- reasonable constraints on weights are
  - at least one score is > 0:
    - if none are, then maximal scoring paths have score 0 & are trivial;
  - expected score is < 0:
    - if ≥ 0, then maximal scoring paths in random seqs will tend to extend through entire sequence
      - more suitable for 'global' than 'local' analyses
- above constraints ⇔ can assume weights are scaled LLRs (will show later)

- Can *choose* prob dist'ns *P*, *Q*, to optimize discrimination of regions to be detected (*like* an LLR test):
  - *P* corresponds to backgd dist'n
    - *sequence graph*: average composition of sequences being scanned
    - *pairwise alignment*: random pairs of residues
  - *Q* corresponds to target dist'n
    - *sequence graph*: composition of regions to be detected – e.g. to detect hydrophobic regions in protein, use residue freqs in observed hydrophobic regions
    - *pairwise alignment*: homologous residue pairs in evolutionarily related sequences

Example where LLR weights *aren't* a natural choice: quality trimming of sequencing reads

- assume have error probs for base calls:
  - e_i = error prob for i-th base call in read, $1 \leq i \leq N$ where N = read length
- want to trim read to that part having error rate ≤ a specified target rate
  - e.g. .05
- construct linked-list directed graph with N edges, & set
  $$w_i = .05 - e_i$$
  as weight on i-th edge
- highest weight path in graph has property that any segment extending path has negative score
  - i.e. avg error rate in extension > .05.


extension must have
neg score
maximum-scoring
segment
extension must have
neg score

Scores on Probability Spaces

- A *scoring system* on a prob space (S,P) is function $W: S \to \mathbf{R}$ ($\mathbf{R}$ = real numbers).
  - $W(s)$ is called the *score* (or *weight*) of s.
- Example: for any prob dist'n $Q \neq P$ on S, the LLR score $W(s) = \log_b(Q(s)/P(s))$.

  This has properties (writing p_s , q_s , w_s for $P(s)$, $Q(s)$, $W(s)$)

1. $w_s > 0$ for at least one s
   - otherwise $q_s \leq p_s$ for all s, and $q_s < p_s$ for at least one s since $Q \neq P$; but then $\sum_s q_s < \sum_s p_s = 1$, so Q is not a probability distribution.
2. $\sum_s p_s w_s < 0$ (by the information inequality)

- above properties also hold for "scaled" LLR $\log_b(q_s/p_s) / \lambda$ where $\lambda > 0$.

- conversely, *any* scoring system W satisfying above two properties is of form $\log_b(q_s/p_s) / \lambda$, for a unique λ and Q

(λ depends on b):

Proof: Take $b = e$ for convenience.

λW is a LLR $\Leftrightarrow e^{\lambda w_s} = q_s / p_s$ for some prob dist'n Q

$\Leftrightarrow \sum_s p_s e^{\lambda w_s} = 1$

$\therefore$ if define

$$f(\lambda) = \sum_s p_s e^{\lambda w_s}$$

then it is enough to show $f(\lambda) = 1$ for a unique $\lambda > 0$, because can then take

$$q_s = p_s e^{\lambda w_s}$$

- $f(\lambda) = 1$ for $\lambda = 0$, $f(\lambda) > 0$ for all λ
- the derivative $f'(\lambda) = \sum_s p_s w_s e^{\lambda\, w_s}$, so $f'(0) = \sum p_s\, w_s < 0$,

  i.e. f decreasing at 0
- $\therefore\ \exists\ \mu > 0$ with $f(\mu) < f(0) = 1$
- $f(\lambda)\to\infty$ as $\lambda\to\infty$ since by assumption some $w_s > 0$
- $\therefore\ f(\lambda) = 1$ for some $\lambda > \mu > 0$
- f is convex
  - i.e. for any λ_1 and λ_2, line segment from the point $(\lambda_1, f(\lambda_1))$ to $(\lambda_2, f(\lambda_2))$ lies above graph of $f(\lambda)$

  since its terms $p_s e^{\lambda\, w_s}$ are convex,
- $\therefore\ \exists$ at most one $\lambda > 0$ with $f(\lambda) = 1$
  - otherwise graph would have ≥ 3 points on line $y = 1$

– this completes the proof.

Karlin-Altschul theory (cont'd)

- *expected* # of maximal segments with scores ≥ a, in 'bkgd' sequence of length N is

$$NKe^{-\lambda a}$$

- where λ, K are constants depending on scoring system
  - λ (as discussed previously) rescales scores to be LLRs
- method assumes sequence is very long
  - i.e. doesn't allow for "edge effects"

Intuition (not a proof!) for K-A formula

- Consider the space of sequences of a *fixed length* $n \leq N$
  - (think of these as the possible subsequences of length n starting at a particular location within a larger sequence of length N.
- Assume LLR scoring system ($\lambda = 1$):
  - score(s) = log(Q(s) / P(s)), for any sequence s of length n, where
    - P = backgd dist'n
    - Q = target dist'n

Intuition cont'd

- What is the total probability of all sequences of score $\geq a$?

$$\log(Q(s) / P(s)) \geq a$$

$$\Rightarrow \quad Q(s) / P(s) \geq e^{a}$$

$$\Rightarrow \quad P(s) \leq e^{-a} Q(s)$$

Summing over all such s:

$$\Sigma_s P(s) \leq e^{-a} \Sigma_s Q(s) = k\, e^{-a} = k\, e^{-\lambda a}$$

for some $k \leq 1$

Intuition cont'd

- (Very) roughly speaking, averaging over possible sequence lengths $n \leq N$, and summing over the N possible start points within a sequence of length N, get $NKe^{-\lambda a}$

- A better (but still incomplete) argument is given in the following slides.

Scores on Probability Spaces (cont'd)

- convenient to
  - assume W takes on integral values
    - rescale and round
      - (loss of precision can be made as small as desired by taking scaling factor large enough);
  - replace original prob space by one induced on the integers by the random variable W – so
    - the sample points are integers
    - prob associated to the integer k is $\sum_{s:w_s=k} p_s$
    - the weight function is now the identity
      - i.e. weight associated to k is k.

Maximal Segments

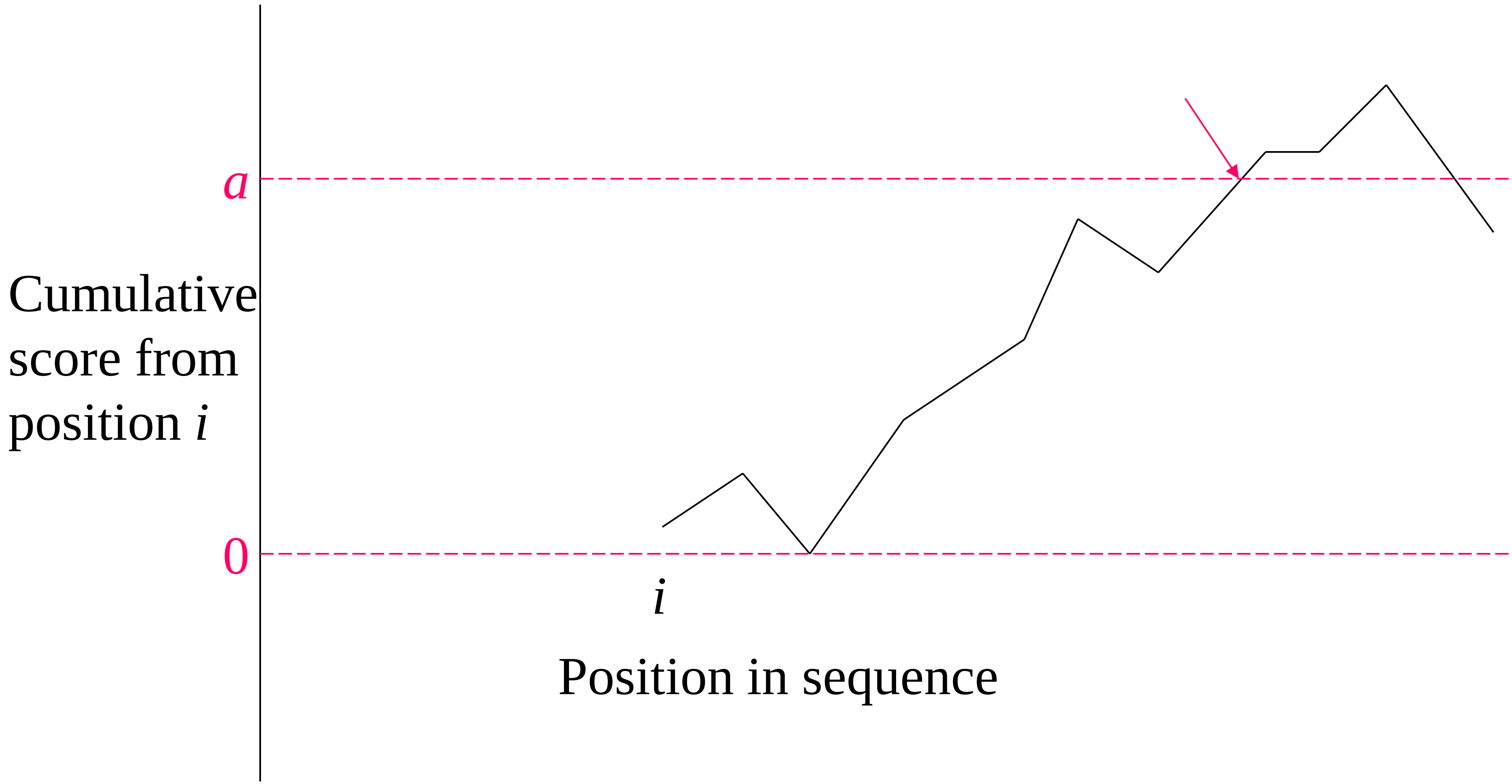
- want prob that maximal segment of score $\geq a$ starts at position i
- this requires two *independent* events to occur:

1. cumulative score
   - starting from value of 0 and
   - adding successive scores while moving to the right from pos'n i,

   must reach value $\geq a$ before reaching value < 0.

   Call prob of this P_1

Moving to right, cumulative score reaches ≥ a before negative value



2. for any $j < i$, score of segment from j to $i - 1$ is < 0
Equivalently,

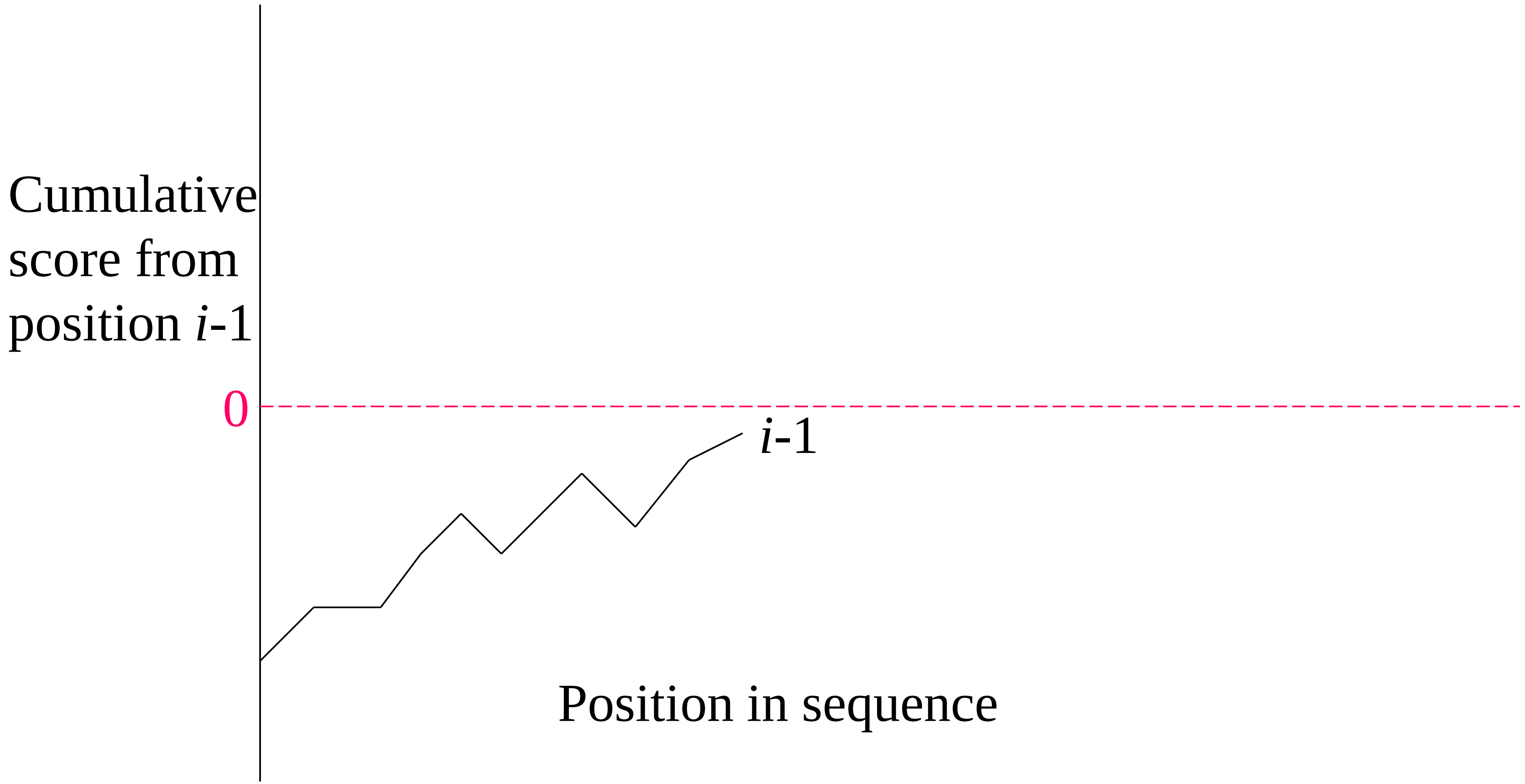
- starting from score 0 and
- adding successive scores while moving to *left* from pos'n $i - 1$
- (and not resetting neg scores to 0)

the score remains < 0. This requires that

- the score k at position $i - 1$ is negative
- cumulative score moving from $i - 1$ leftward never gets back to 0 from k

Call prob of this P_2

Moving to left, cumul score always < 0



Analogy to random walk/gambler's ruin

- cumulative score, counting from particular position in sequence, corresponds to
  - total distance walked, or
  - gambler's net worth
- with each step having probability p_k of moving distance k
  - k positive $\Rightarrow$ forwards
  - k negative $\Rightarrow$ backwards
- *stop* when reach
  - value < 0 (out of money!); or
  - value $\geq a$

"random walk with absorbing barriers at 0 and a"

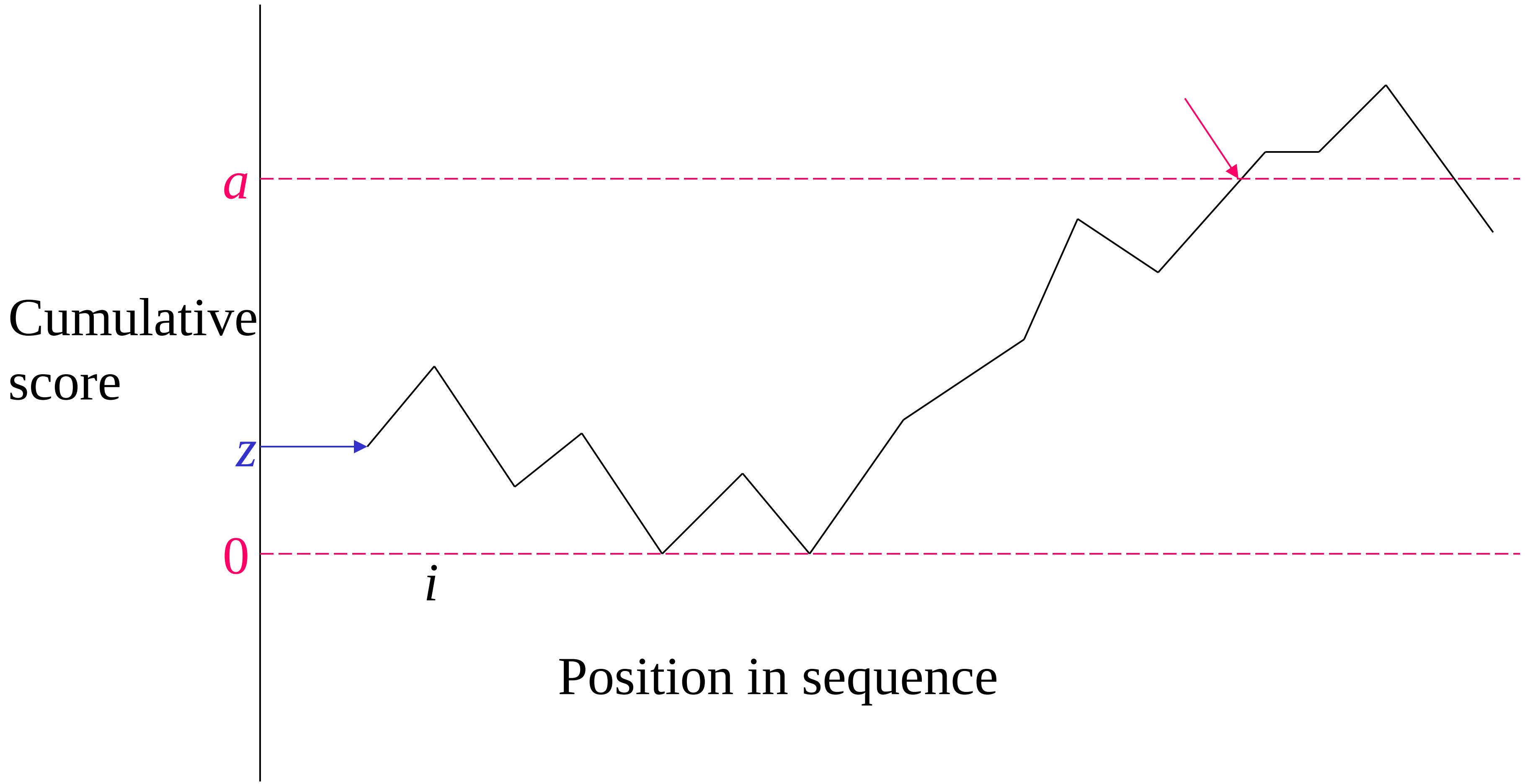
- estimate P_1 and P_2 and *multiply* (since cond’ns are independent) to get

  prob (max segment of score $\geq a$ starts at i)

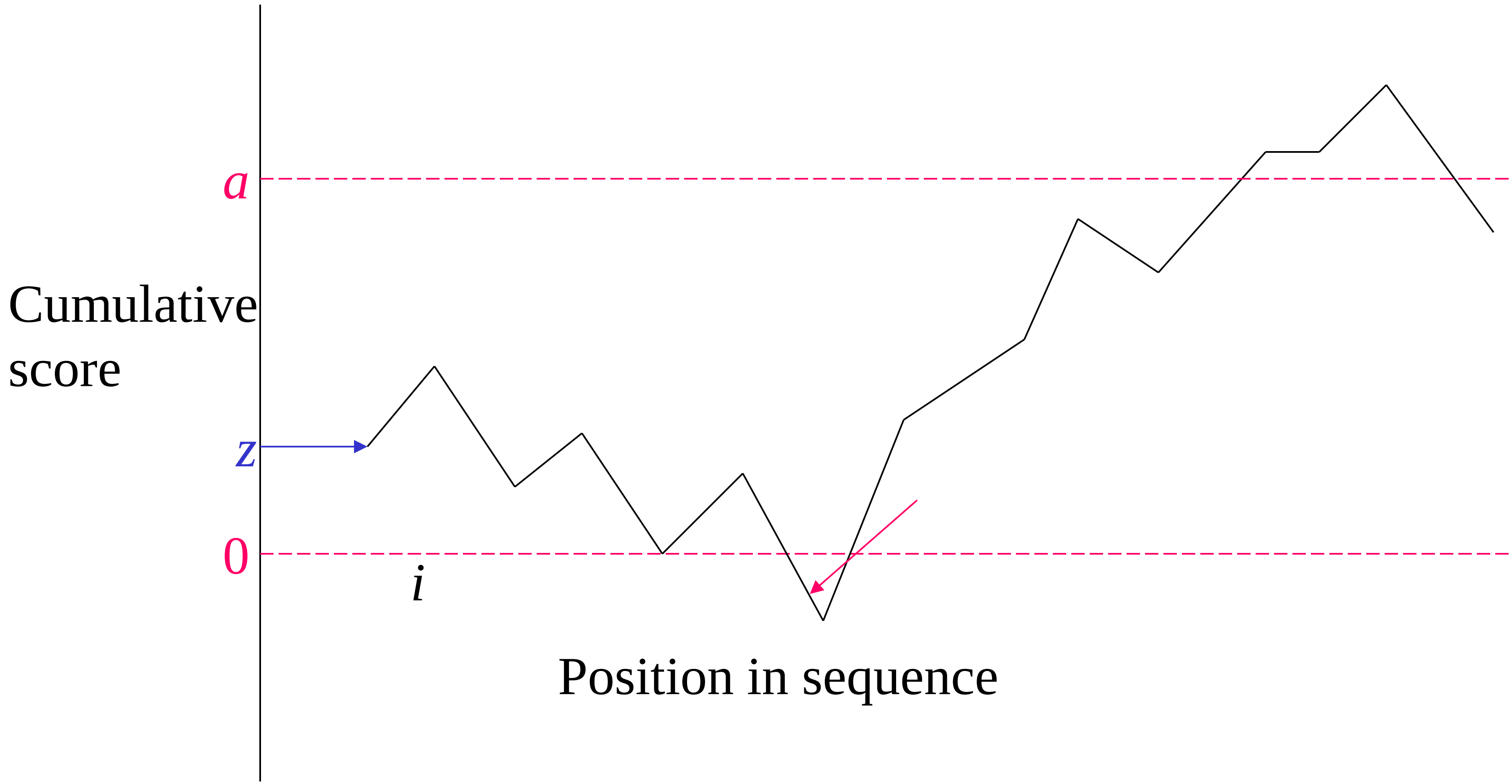
Estimating P_1

- consider a more general situation:
  - assume start with score = z (an integer) instead of 0,
  - again consider cum score moving to right from position i
  - what is prob u_z of getting to target score $\geq a$ before getting to < 0?
- $P_1 = u_0$

Success (Reach ≥ a First)



Failure (Reach < 0 First)



Non-rigorous derivation

- intuition (*not a proof*!) for why P_1 should be approximately $e^{-\lambda a}$:

  for any $a > b$, let

  $\mathsf{P}(a \mid b)$ = prob that, starting from cumul score = b, eventually reach cumul score a

  - (*ignoring* whether drop below 0 first – which is one reason why this isn't a proof!)

  Then

  - $\mathsf{P}(a \mid b) = \mathsf{P}(a - b \mid 0)$
  - $\mathsf{P}(a + a' \mid 0) = \mathsf{P}(a' \mid 0)\,\mathsf{P}(a + a' \mid a') = \mathsf{P}(a' \mid 0)\,\mathsf{P}(a \mid 0)$

$\therefore$ the function $a \to \mathsf{P}\,(a \mid 0)$

takes sums to products

$\therefore$ $\mathsf{P}\,(a \mid 0) = e^{-\mu a}$ for some μ

What is μ?

Consider first step, starting at 0:

prob it has size k is p_k

Considering all possible sizes of 1st step:

$\mathsf{P}(a \mid 0) = \sum_k p_k \, \mathsf{P}(a \mid k) = \sum_k p_k \, \mathsf{P}(a - k \mid 0)$

$\Rightarrow e^{-\mu a} = \sum_k p_k \, e^{-\mu (a-k)}$

$\Rightarrow$ (cancelling $e^{-\mu a}$) $1 = \sum_k p_k \, e^{\mu k}$

$\Rightarrow \mu = \lambda$ (by definition of λ)

$\Rightarrow \mathsf{P}(a \mid 0) = e^{-\lambda a}$