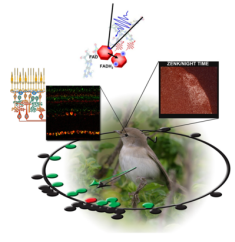
Rare-event simulations for score statistics of multiple sequence alignments

Pascal Fieth, Alexander K. Hartmann

Institut für Physik
Universität Oldenburg

Leipzig, 28.11.2014



Outline

Biological Background

Sequence Similarities

- Relations, ...
- Functional groups, protein structure...

Evolution of DNA sequences

Copies by DNA polymerase:

- Inherit = match
- Replace = mismatch
- Insert/delete = gap

A G C T A

A T T A

DNA alignment

Biological Background

Sequence Similarities

- Relations, ...
- Functional groups, protein structure...

Evolution of DNA sequences

Copies by DNA polymerase:

- Inherit = match
- Replace = mismatch
- Insert/delete = gap

DNA alignment

Biological Background

Sequence Similarities

- Relations, ...
- Functional groups, protein structure...

Evolution of DNA sequences

Copies by DNA polymerase:

- Inherit = match
- Replace = mismatch
- Insert/delete = gap

DNA alignment

Biological Background

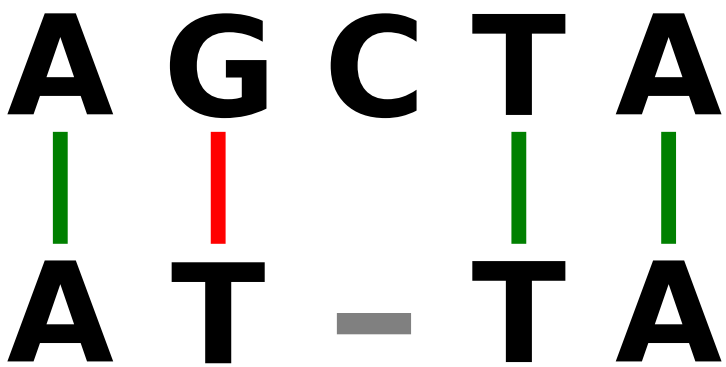
Sequence Similarities

- Relations, ...
- Functional groups, protein structure...

Evolution of DNA sequences

Copies by DNA polymerase:

- Inherit = match
- Replace = mismatch
- Insert/delete = gap



DNA alignment

$$S = \sum_{\text{aligned},k} s(x_{ik}, y_{jk}) - \sum_{\text{gaps},g} d + (l_g - 1)e$$

Scoring of Residue Pairs

Biologists provide *substitution matrices*
Reflect statistical model of substitutions

	A	⋯	W	Y	V
A	4	⋯	-3	-2	0
⋮	⋮	⋱			⋮
W	-3		11	2	-3
Y	-2		2	7	-1
V	0	⋯	-3	-1	4

BLOSUM62: extract

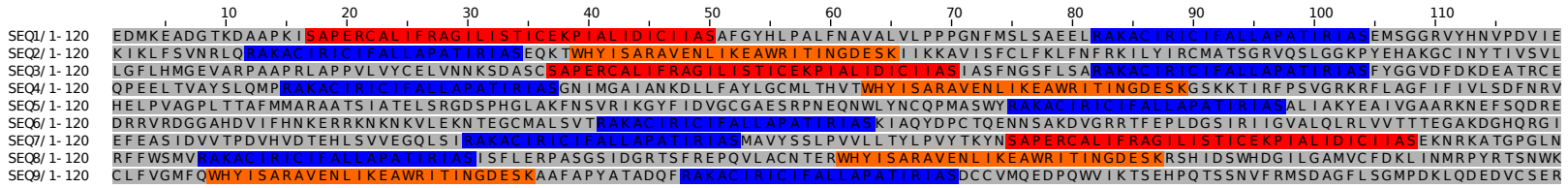
Optimal Pairwise Alignment

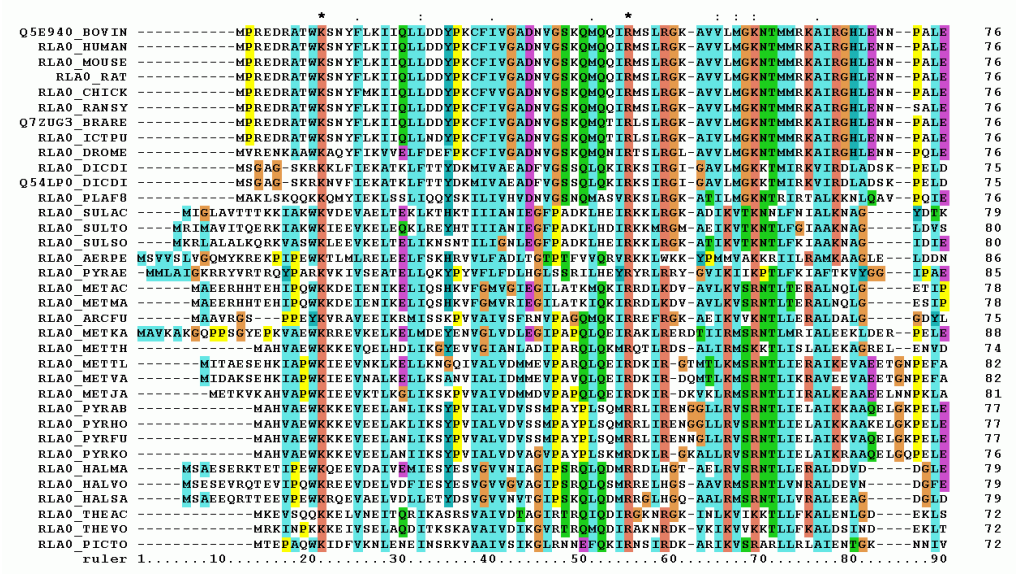
- Many possible alignments
- Find highest scoring one $\rightarrow$ most likely actual relation

$\Rightarrow$ Optimisation problem, solve by *dynamic programming* $\mathcal{O}(L^2)$
[Smith,Waterman;J.Mol.Bio;1970]

Multiple Sequence Alignments

- Find the optimal alignment of $N > 2$ sequences
- Local alignment: best scoring mutual subsequence
- Dynamic programming in $\mathcal{O}(L^N)$ for sequences of length L





Local MSA

Global MSA of protein P0 in several organisms

Miguel Andrade, http://en.wikipedia.org/wiki/File:RPLP0_90_ClustalW_aln.gif

Rare Event Simulation

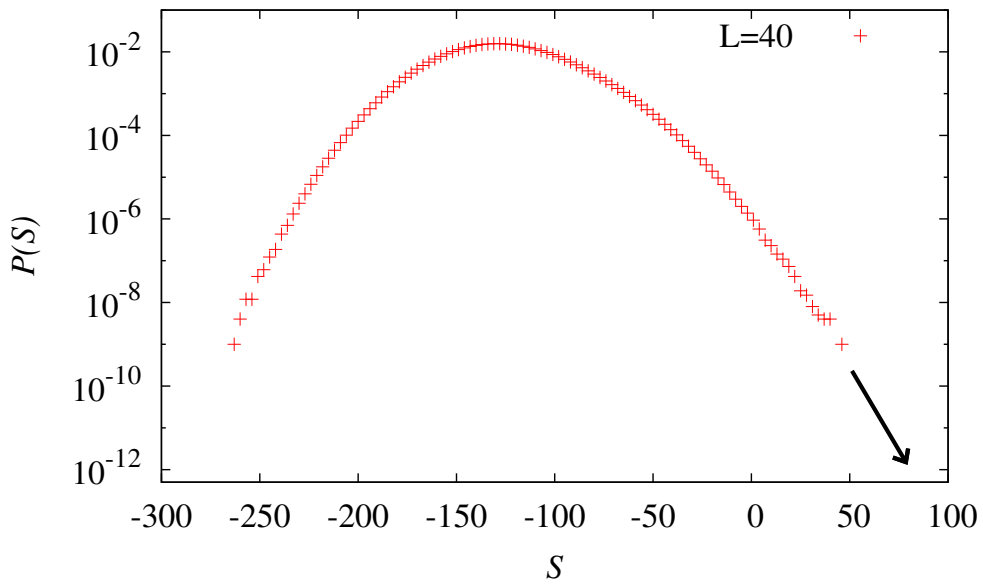
Sequence Similarity

- Score so far only indicator for alignment optimization
- Relation or random agreements? How ”good” is the obtained score S_{obs}?

$\Rightarrow$ Score statistics as criterium: $P(S \geq S_{\text{obs}})$

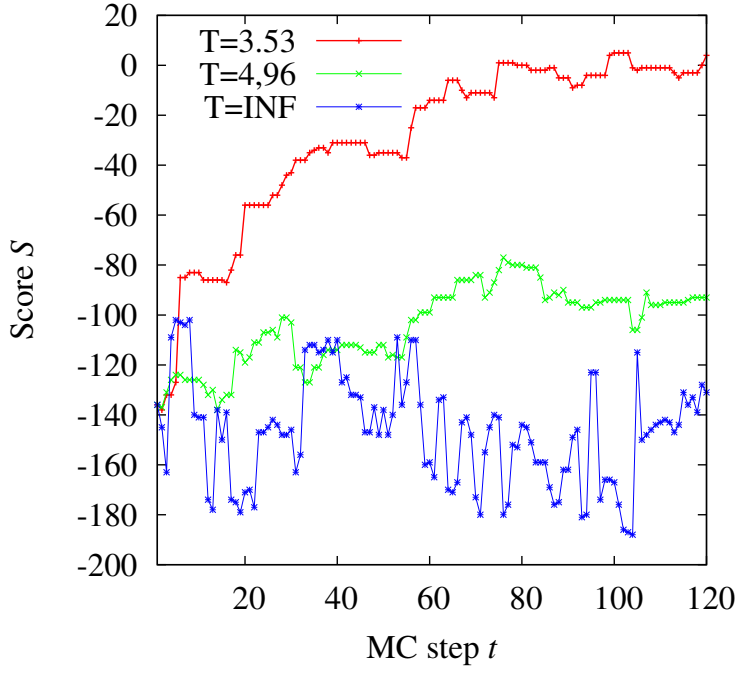
Simple Sampling

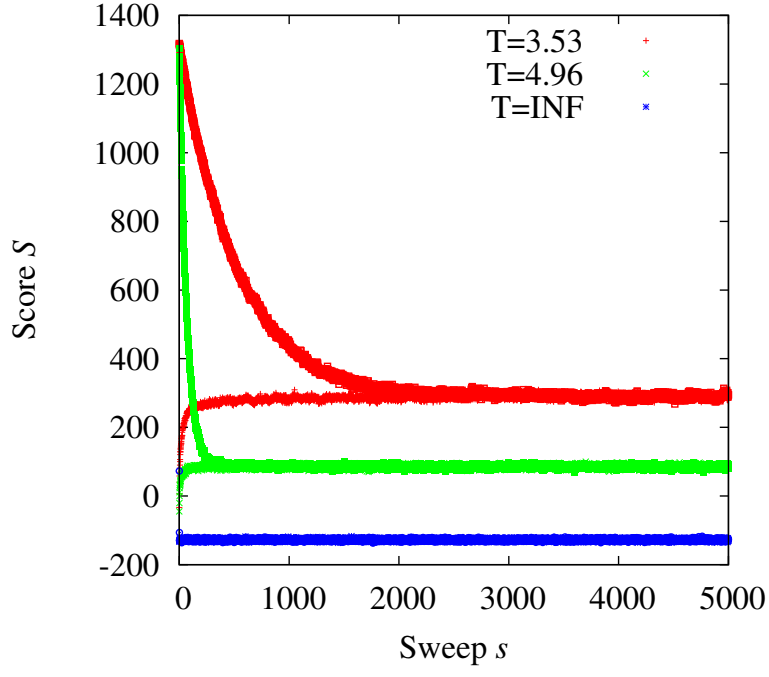
- Only covers high probability region
- Biologically relevant: low probability, high scoring tail



Mapping to Statistical Mechanics [Hartmann;Phys.Rev.E;2002]

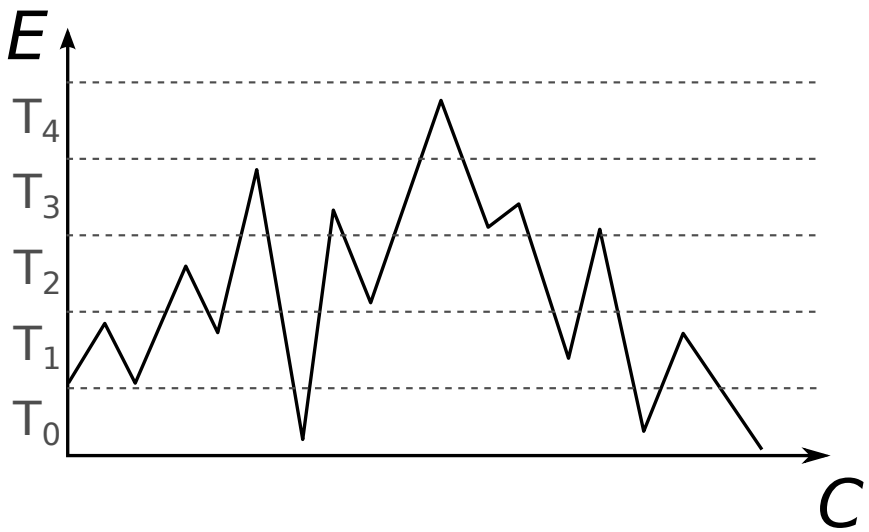
- Decrease energy $\rightarrow$ increase score: Use $E = -S$
- Simulate Markov Chain of sequences at finite "temperature" with probabilities $P(S) \cdot \exp(\frac{S}{T})$.

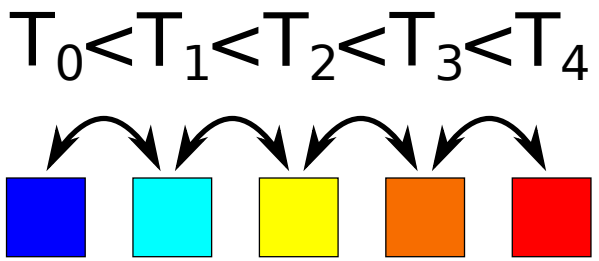




Parallel Tempering

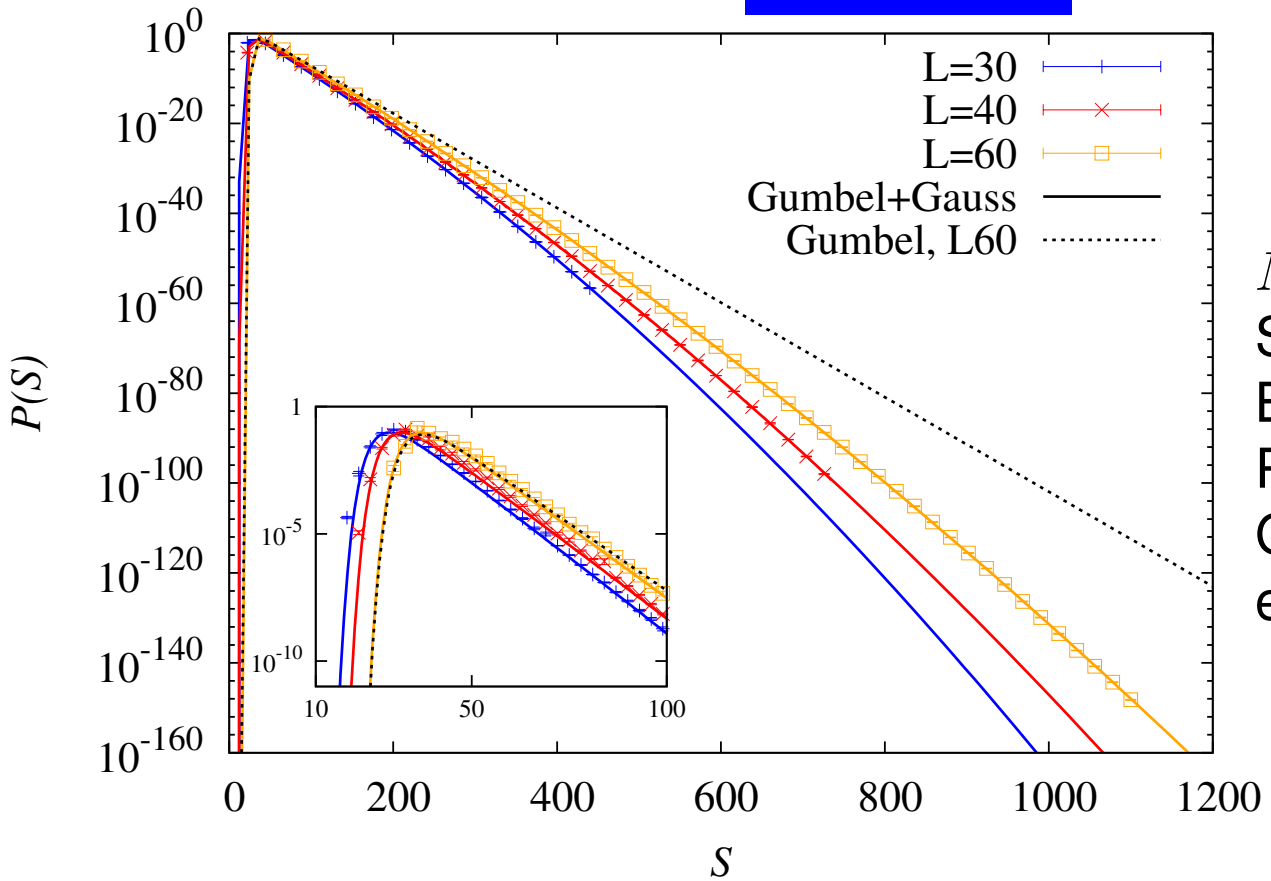
- To cover large score range: simulate at different temperatures
- After one sweep switch systems of neighboring temperatures with probability
  $P_{\text{sw}}(T_i) = \max(1, \exp(\Delta S_i \Delta \beta_i))$
  with $\Delta S_i = (S_i - S_{i+1})$,$\Delta \beta_i = \frac{1}{T_i} - \frac{1}{T_{i+1}}$
- Avoid trapping in local maxima





- Parallel simulation using MPI: 15 to 20 temperatures

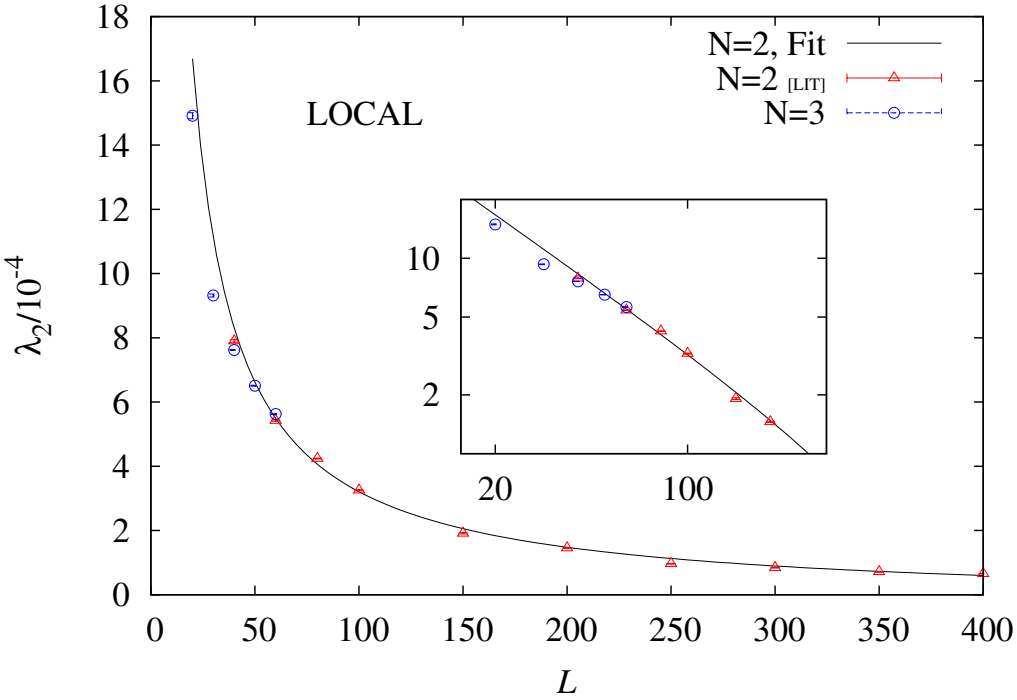
Results



$N = 3$
Substitution matrix: BLOSUM62
Penalties:
Gap open: $d = 12$
extend: $e = 1$

- Previous studies found Gumbel distribution with Gaussian correction for $N = 2$ sequences: [Wolfsheimer,Burghardt,AKH;Alg.Mol.Bio.;2007]

$$P(S) \propto \lambda \exp\left(-\lambda(S - S_0) - e^{-\lambda(S-S_0)}\right) \exp\left(-\lambda_2(S - S_0)^2\right)$$

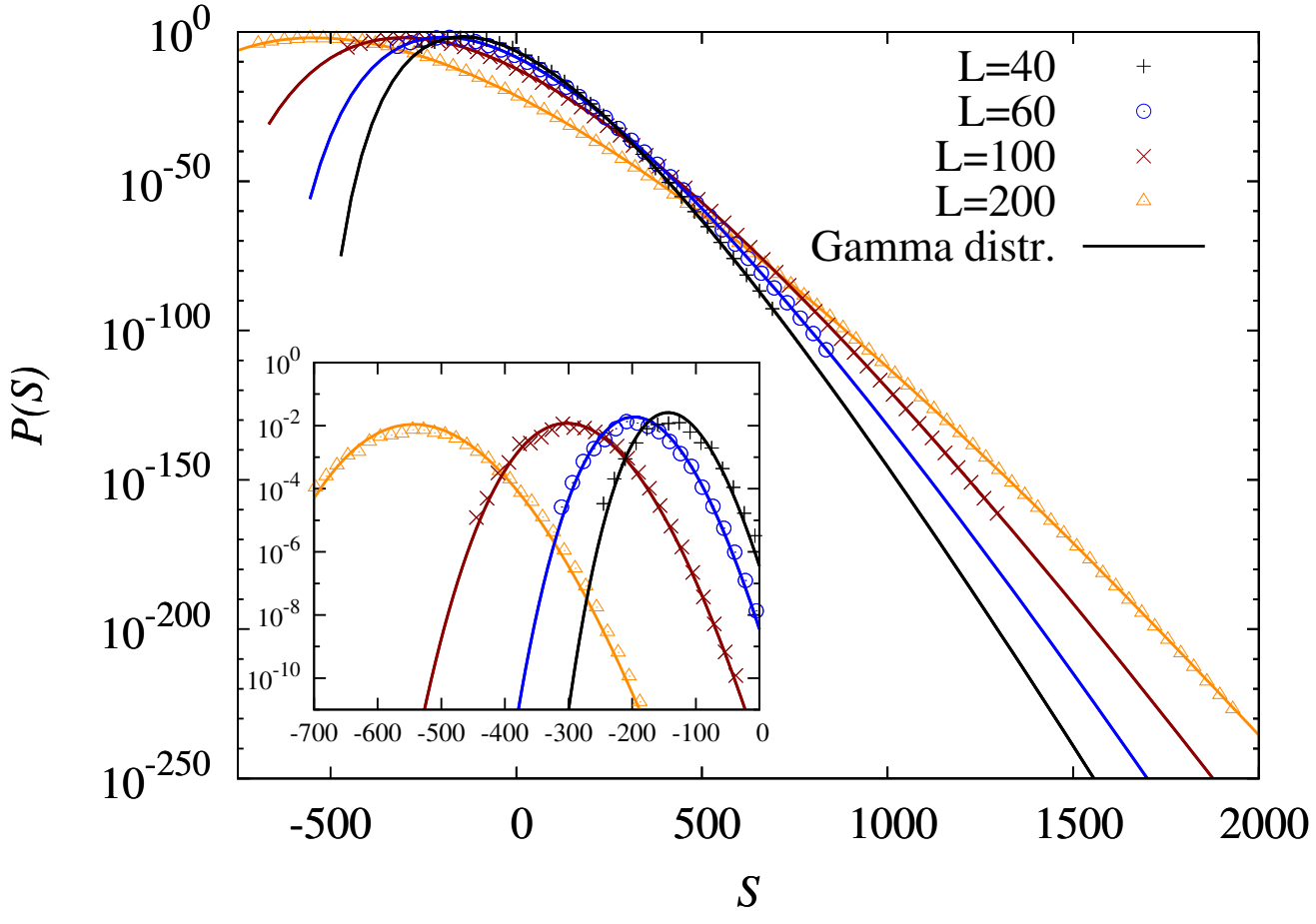


Length-Dependence of Gaussian Correction

- Use score per pair $\frac{S}{(N-1)(N-2)}$
- Calculate λ_2

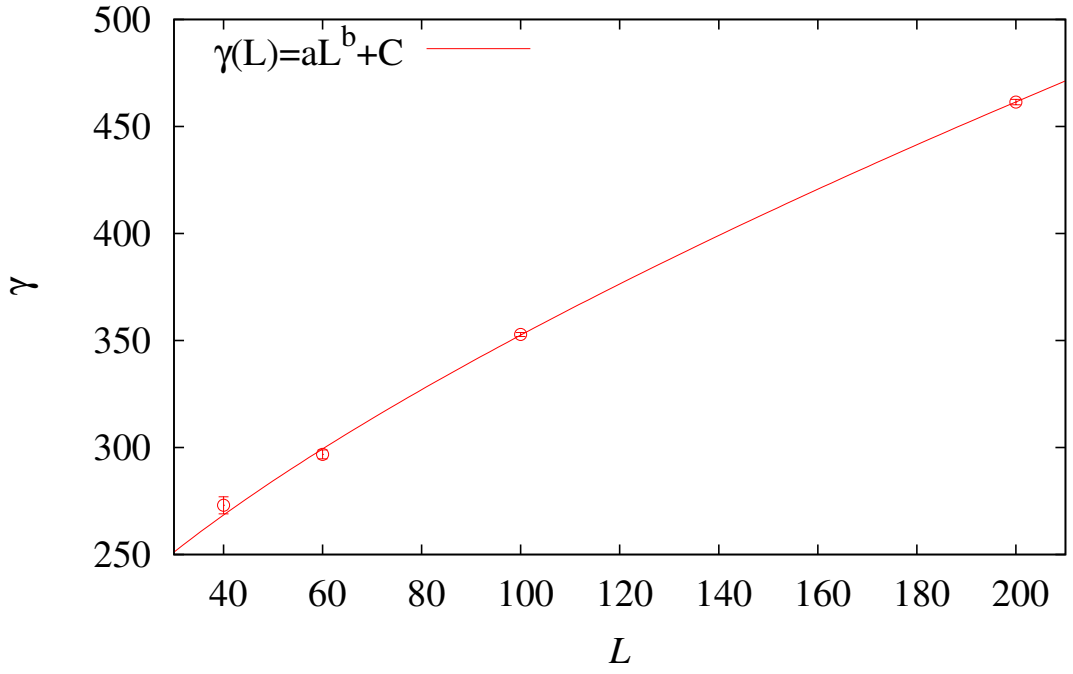
$\Rightarrow$ decreases with increasing sequence length
$\Rightarrow$ coincides so far with findings for $N = 2$

[Wolfsheimer,Burghardt,Hartmann;Alg. Mol. Bio.;2007]



Global MSA, $N = 3$, Gamma-distribution
[Pang,Tang,Chen,Tao; BMC Bioinf; 2005]

$$P_\Gamma(S) = \frac{\lambda^\gamma}{\Gamma(\gamma)}(S-\mu)^{\gamma-1}\exp\left(-\lambda(S-\mu)\right)$$



Length-Dependence of γ

- γ increases with growing sequence length
- $\gamma = 1$: exponential distribution
- $\gamma \to \infty$: normal distribution

Conclusion

Large Deviation Simulation

- Mapping to statistical mechanics
- Successful approach to sample small probabilities

Local Alignments

- Behavior for $N = 2$ confirmed for $N = 3$
- Gumbel distribution of scores with Gaussian correction in tail
- Deviation from the standard Gumbel distribution
  e.g. relevant in database searches

  [Wolfsheimer,Herms,Rahmann,Hartmann;BMC Bioninformatics;2011]

Global Alignments

- No analytical solution
- Gamma distribution better candidate than Gumbel distribution

Prospects

- Check if this holds for other substitution matrices
- Analyse dependence on gap penalties
- Develop heuristic for gapped local MSA
- Check results for sequence sets with $N > 3$

Global Alignments

- No analytical solution
- Gamma distribution better candidate than Gumbel distribution

Prospects

- Check if this holds for other substitution matrices
- Analyse dependence on gap penalties
- Develop heuristic for gapped local MSA
- Check results for sequence sets with $N > 3$

Thank you for your attention!

Dynamic Programming

		H	E	A	G	A	W
	0	← **-8**	← -16	← -24	← -32	← -40	← -48
		↖	↖	↖		↖	
P	-8	-2	-9	**-17**	← **-25**	-33	← -41
	↑	↑	↖	↖		↖	
A	-16	-10	-3	-4	← -12	**-20**	← -28
	↑	↑	↑	↖	↖	↖	↖
W	-24	-18	-11	-6	-7	-15	**-5**
	↑	↖	↖	↖	↖	↖	↑
H	-32	-14	-18	-13	-8	-9	**-13**

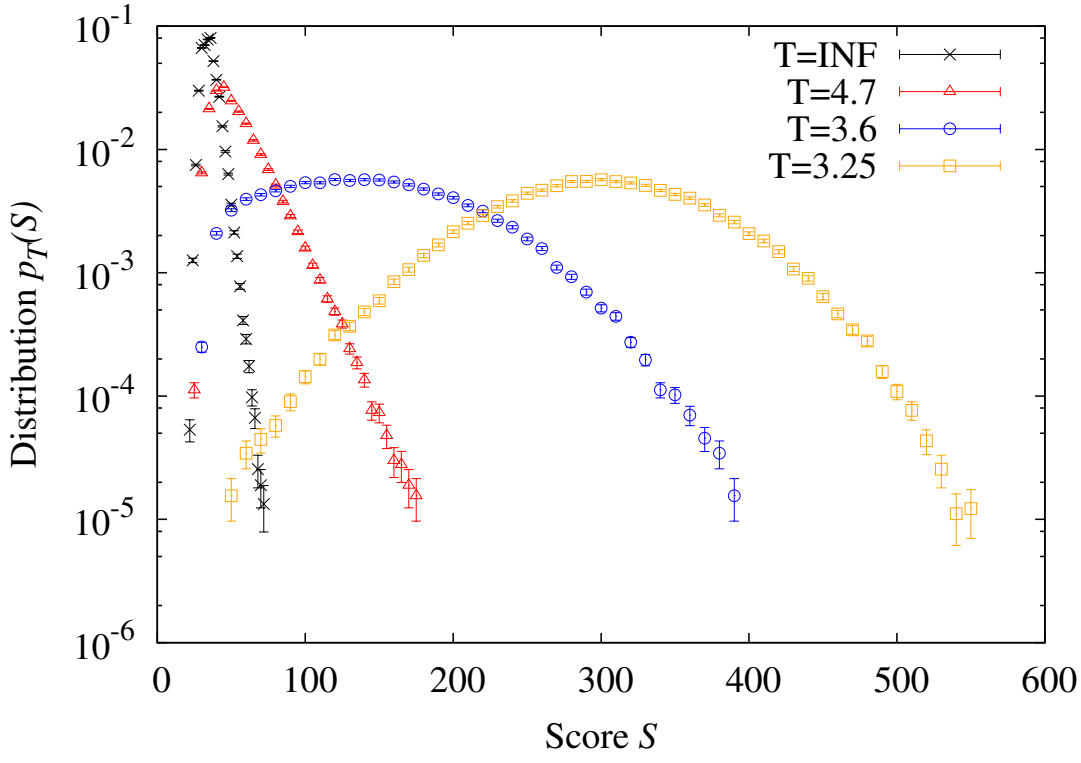
Dynamic Programming matrix for fixed gap costs [Durbin1998]

H E A G A W -
- - P - A W H

$$P(S) = p_T(S) Z_T \exp(-S/T)$$

Obtaining $P(S)$

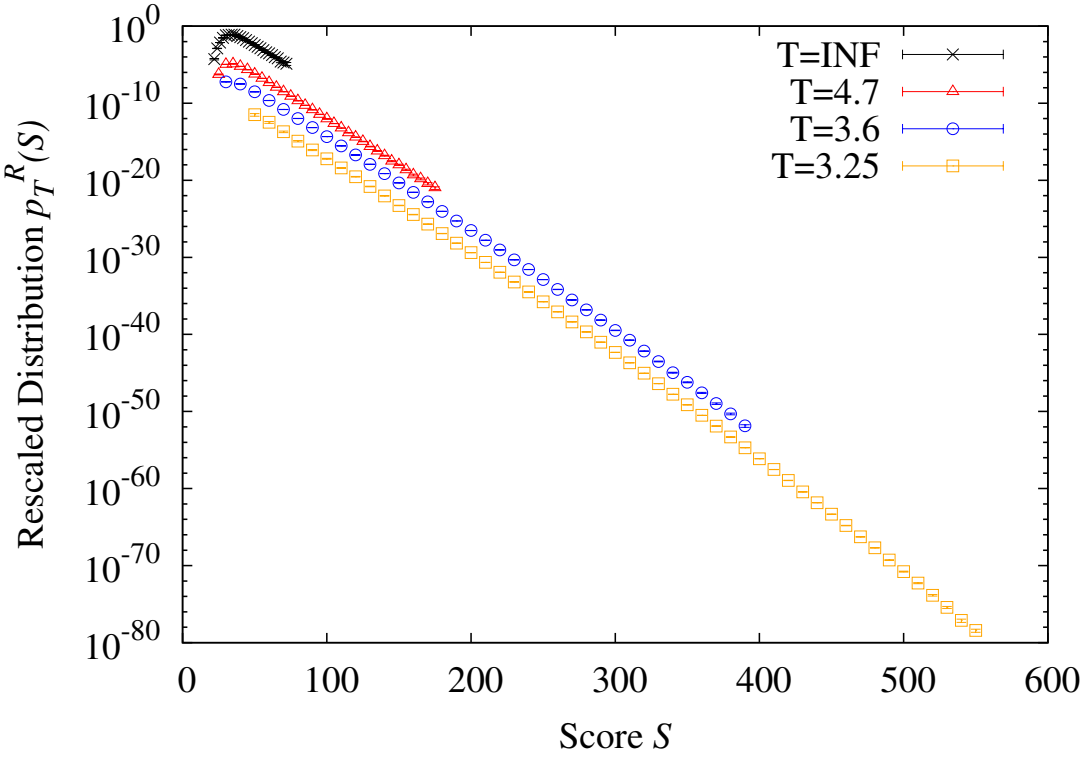
- Simulations give scaled distributions $p_T(S)$



$$P(S) = p_T(S) Z_T \exp(-S/T)$$

Obtaining $P(S)$

- Simulations give scaled distributions $p_T(S)$
- Rescale with $\exp(-S/T)$



$$P(S) = p_T(S) Z_T \exp(-S/T)$$

Obtaining $P(S)$

- Simulations give scaled distributions $p_T(S)$
- Rescale with $\exp(-S/T)$
- Assume $P(S) = p_\infty(S)$
- $\rightarrow$ shift on logarithmic scale