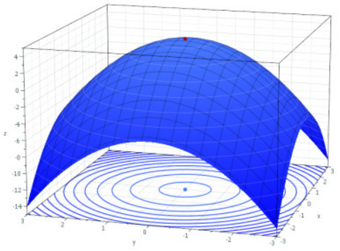
Gibbs Sampling

Stephen F. Altschul

National Center for Biotechnology Information
National Library of Medicine
National Institutes of Health

Optimization in High-Dimensional Space

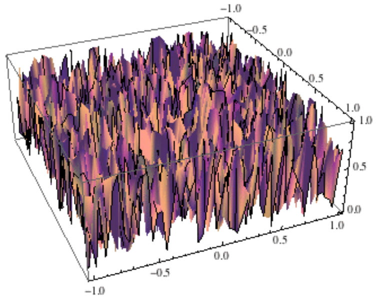
Smooth and simple landscapes



Relatively easy to find optimum.

Algorithms: Newton’s method; gradient descent.

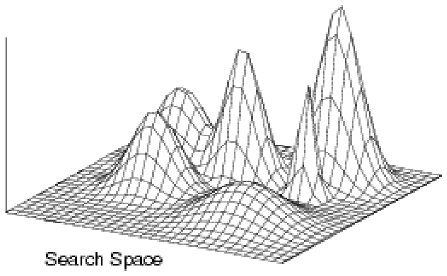
Random landscapes



Finding optimal solution intractable.

Algorithms: Brute force enumeration.

Rough but correlated landscapes



Difficult to find provably optimum solution.

Fairly effective heuristic methods available.

Algorithms: Simulated annealing; Gibbs sampling.

  Success depends on details of landscape.

  Difficulties: Local optima.

Images courtesy of the internet

Local Multiple Alignment

Simple Version of Problem

Input: N sequence; pattern width W.

Problem: Find "highest-scoring", ungapped local multiple alignment, involving one segment of length W from each sequence.

Search space: L^N

One may score the local alignment in various ways. Here, we will use *BILD* scores.

Gibbs sampling:

Given an alignment, one may easily derive a profile or scoring matrix.

Given a profile, one may easily calculate its likelihoods, implied by various segments within a sequence.

Gibbs sampling alternates between generating profiles from given alignments, and sampling alignment positions based on given profile, until "convergence".

Lawrence, C.E., *et al.* (1993) *Science* **262**:208-214.

1. Initialization

Choose random length-*W* segments from within the input sequences:

MTQPSKTTKLTKDEVDRLISDYQTKQDEQAQETL**VRVYTNLVDMLAKKY**SKGKSFHEDLRQVGMIGLLGAIKRYD
PVVGKSFEAFAIPTIIGEIKRFLRDKTWSVHVPRRIKELGPRIKMAVDQLTTETQRSPKVEEIAEFLDVSEEEVL
ETMEMGKSYQALSVDHSIEADSDGSTVTILDIVGSQEDGYERVNQQLMLQSVLHVLSDREKQIIDLTYIQNKSQK
ETGDILGISQMHVSRLQRKAVKKLREALIEDPSMELM

MPPLFVMNNEILMHLRALKKTKKDVSLHDPIGQDKEGNEISLIDVLKSENEDVIDTIQLNMELEKVKQYIDILDD
REKEVIVGRFGLDLKKEKTQREIAKELGISRSYVSRIEKR**ALMKMFHEFYRAEKE**KRKKAKGK

MELRDLDLNLLVVFNQLLVDRRVSITAENLGLTQPAVSNALKRLRTSLQDPLFVRTHQGMEPTPYAAHLAEPVTS
AMHALRNALQHHESFDPLTSERTFTLAM**TDIGEIYFMPRLMDV**LAHQAPNCVISTVRDSSMSLMQALQNGTVDLA
VGLLPNLQTGFFQRRLLQNHYVCLCRKDHPVTREPLTLERFCSYGHVRVIAAGTGHGEVDTYMTRVGIRRDIRLE
VPHFAAVGHILQRTDLLATVPIRLADCCVEPFGLSALPHPVVLPEIAINMFWHAKYHKDLANIWLRQLMFDLFTD

MNAYTVSRLALDAGVSVHIVRDYLLRGLLRPVACTTGGYGLFDDAALQRLCFVRAAFEAGIGLGALARLCRALDA
ANCDETAAQ**LAVLRQFVERRREAL**ANLEVQLAAMPTAPAQHAESLP

⋮

2. Remove one segment from alignment

Select a sequence “*X*” at random from among the input sequences, and remove its segment from the multiple alignment:

VRVYTNLVDMLAKKY
ALMKMFHEFYRAEKE
TDIGEIYFMPRLMDV
LAVLRQFVERRREAL
PSPLYPWMRSQFGKC
DDTAIRTVLNQALSR
QWERGDSEPTGKNLF
YHHIKKEKSPKGKSS
RIESALLNKIAMLGT

3. Construct a profile from the remaining alignment

Multiple alignment:

VRVYTNLVDMLAKKY
ALMKMFHEFYRAEKE
TDIGEIYFMPRLMDV
LAVLRQFVERRREAL
⋮

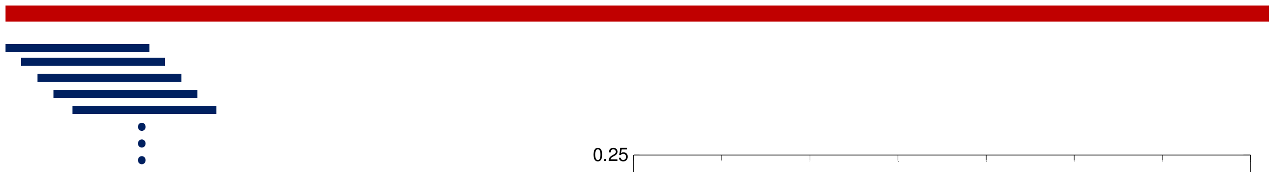
⇓ ⇓

Log-odd scores:

A:	1	1	0	−1	0	
C:	−2	0	−2	−1	−1	…
D:	0	2	−3	0	1	
E:	1	1	−2	0	0	
⋮			⋮			

4. Calculate relative likelihoods at all positions, and sample

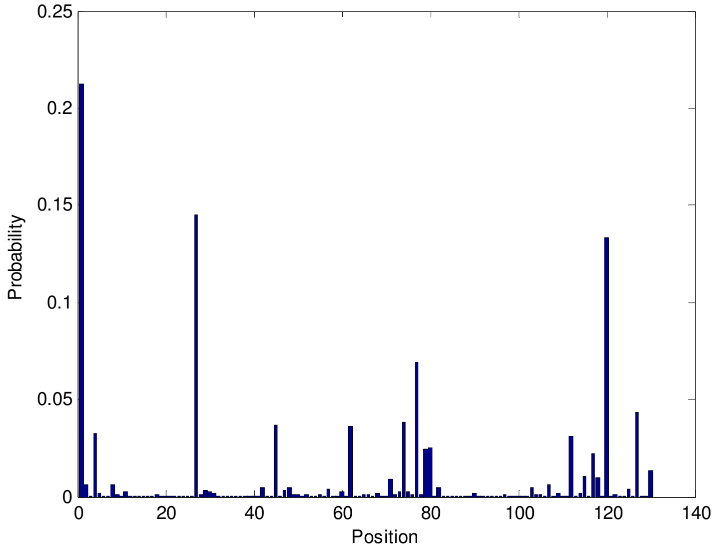
Sequence X:



⇒

Normalized likelihoods:

$(2^{bit\ score})$



Sample a random position from sequence X, weighted by normalized likelihoods.

Add the segment at this position to the multiple local alignment.

If this new alignment is better than any so far seen, remember it.

If there has been no improvement in the last I iterations, stop.

Otherwise, return to step 2, and remove a new segment.

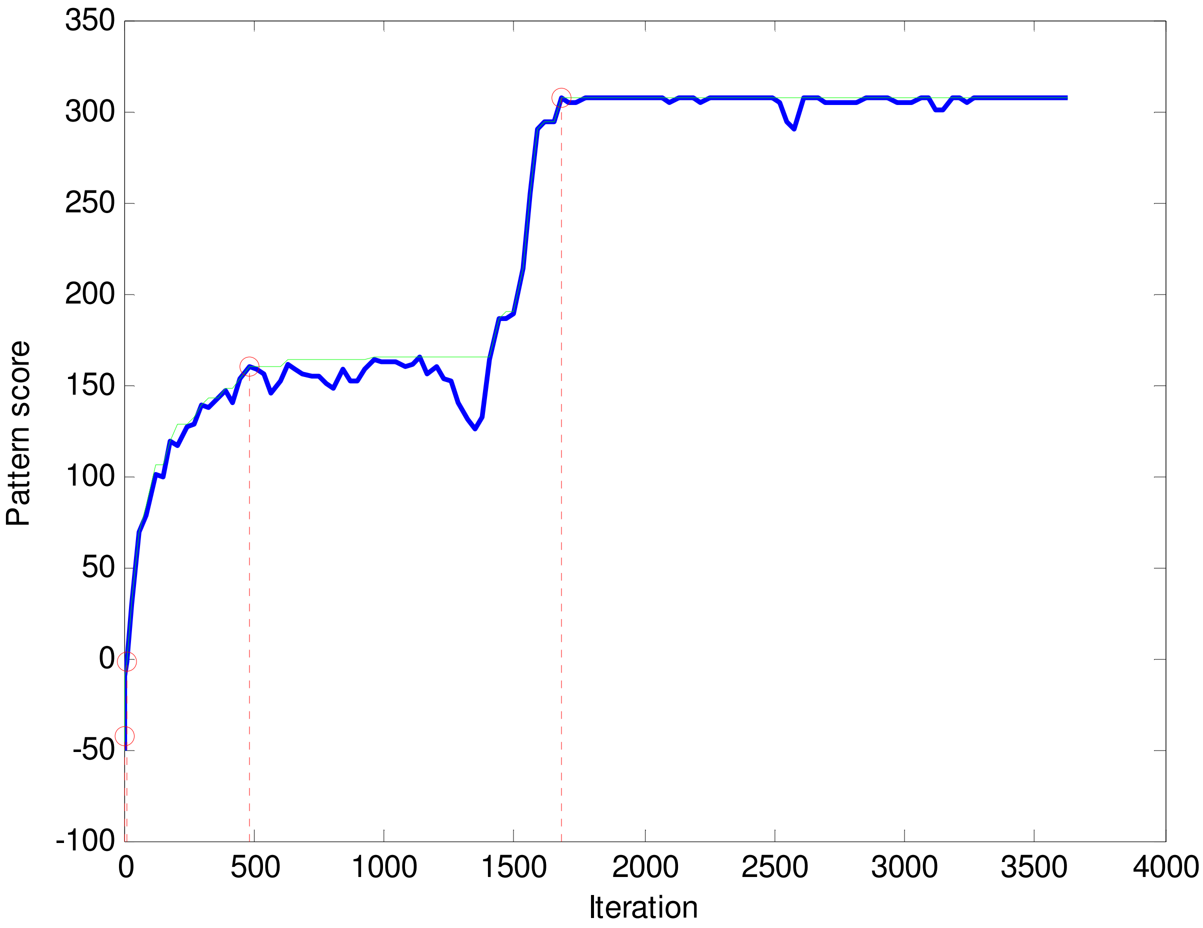
Why Does the Algorithm Work?

When no common pattern is represented in the multiple alignment, the positions in sequence *X* to be sampled have roughly equal likelihoods, so the algorithm performs a random walk through the solution space.

Once a single segment is chosen that is similar to segments found in most or all sequences, these other segments are slightly favored, and a second related segment may well be sampled.

As more related segments are found, the process accelerates, converging on a locally optimal solution. If there are no other good local optima, this solution has a good chance or being the global optimum.

Behavior of the Objective Function



Input
30 HTH sequences
Pattern width 15

Search space
~10^{68}

Time
< 1.0 seconds

The Evolving Multiple Alignment

MTQPSKTTKLTKDEV	SKTTKLTKDEVDRLI	GISQMHVSRLQRKAV	ETGDILGISQMHVSR
MPPLFVMNNEILMHL	FVMNNEILMHLRALK	GISRSYVSRIEKRAL	EIAKELGISRSYVSR
VVFNQLLVDRRVSIT	QLLVDRRVSITAENL	TVRDSSMSLMQALQN	ITAENLGLTQPAVSN
WFQNRRMKWKKENKT	RYLTRRRRIEIAHAL	GVPQQQQQQQQQPSQ	EIAHALCLTERQIKI
SGTGKELVARALHDY	KELVARALHDYGRRR	KLDAQALERLKQHRW	RAADLLGLNRNTLRK
RIRYRRKNLKHTQRS	RRKNLKHTQRSLAKA	PESEQDTQLAEMRAR	SLAKALKISHVSVSQ
ALDAGVSVHIVRDYL	GVSVHIVRDYLLRGL	VLRQFVERRREALAN	RAAFEAGIGLGALAR
QLNGQDVNDLYELVL	QDVNDLYELVLAEVE	PLRDSVKQALKNYFA	RAALMMGINRGTLRK
LEIYHHIKKEKSPKG	HHIKKEKSPKGKSSI	FIMESNLTKVEQHTL	EVAKKCGITPLQVRV
SQISRWKRDWIPKFS	RWKRDWIPKFSMLLA	GVDKSQISRWKRDWI	KTAEAVGVDKSQISR
GSVAVLIKDEEGKEM	VLIKDEEGKEMILSY	RIAQTLLNLAKQPDA	EIGQIVGCSRETVGR
TINADGSVYAEEVKP	QTKTAKDLGVYQSAI	GVYQSAINKAIHAGR	KTAKDLGVYQSAINK
EIVTAGALKYQENAY	AGALKYQENAYRQAA	GISDAAVSQWKEVIP	AVAKALGISDAAVSQ
QLLLRRMEAINESLH	RRMEAINESLHPPMD	LLEQLLLRRMEAINE	SVAQHVCLSPSRLSH
DLSGKMPNLRQQMMR	QDMILLLSKKNAEER	NLRQQMMRLMSGEIK	DIGNYLGLTVETISR
GGLDSYIRAANAWPM	SYIRAANAWPMLSAD	RVRQLEKNAMKKLRA	ELADRYGVSAERVRQ
TRLAWPGNVRQLENT	RLARHFLQIAARELG	MLPDSWATLLGQWAD	EAARLLGWGRNTLTR
ETAATMKDVALKAKV	TMKDVALKAKVSTAT	KVSQATRNRVEKAAR	DVALKAKVSTATVSR
PRSASHYLLSDQKSR	LVEEKRRAAKLAATL	LLSDQKSRLVEEKRR	DAAALLGVSEMTIRR
YHNEQKERQAIEQLI	QKERQAIEQLIRHRC	KERQAIEQLIRHRCA	DVARLAGVSVATVSR
RLLQLSQGQAVKGNQ	AMLVANDQMALGAMR	ALADSLMQLARQVSR	DVAEYAGVSYQTVSR
TRPTEKQYETLENQL	KNKRALLDALAIEML	VLEDQEHQVAKEERE	KLAQKLGVEQPTLYW
SNSLKAAPVELRQWL	KAAPVELRQWLEEVL	YSAAMAEQRHQEWLR	ELKNELGAGIATITR
AFVKFNCAALPDNLL	FNCAALPDNLLESEL	LSRATEASKTLQEVL	KAARLLGMTPRQVAY
EQLNEREKQIMELRF	EREKQIMELRFGLVG	GISQSYISRLEKRII	DVADMMGISQSYISR
EDKISGTKSERPGLK	SGTKSERPGLKKLLR	MERELIVERTKAGLE	KVAIIYDVGVSTLYK
TIHQPKDSLGETAFN	PKDSLGETAFNMLLD	FEPESGYRAMQQILS	DVAKRANVSTTTVSH
FIGGEDEPGKADIRE	EDEPGKADIREVAFA	FSSSSGYELAKQMLA	DIAIEAGVSLATVSR
ARQQEVFDLIRDHIS	EVFDLIRDHISQTGM	HISQTGMPPTRAEIA	EIAQRLGFRSPNAAE
EDEELAELAKKVAHL	LAELAKKVAHLLTKE	GINESQISRWKGDFI	KVADALGINESQISR
1 iteration	10 iterations	480 iterations	1680 iterations

Phase Shifts

The Gibbs sampling algorithm may easily converge on a local optimum that is a "phase-shifted" version of the global optimum. Why?

Optimal solution:
Solution found:

SQKETG**DILGISQMHVSRLQR**KAVKKL
TQREIA**KELGISRSYVSRIEK**RALMKM
RVSITA**ENLGLTQPAVSNALK**RLRTSL
CFVRAA**FEAGIGLGALARLCR**ALDAAN
… RRIEIA**HALCLTERQIKIWFQ**NRRMKW …
NQIRAA**DLLGLNRNTLRKKIR**DLDIQV
TQRSLA**KALKISHVSVSQWER**GDSEPT
EKEEVA**KKCGITPLQVRVWFI**NKRMRS
GTEKTA**EAVGVDKSQISRWKR**DWIPKF

One remedy is to add a separate "phase-shift sampling step".

No segments are removed, but likelihoods are calculated for the current alignment and several phase-shifted alternatives. These alignments are then sampled among.

This can be understood as changing the topology, of definition of distance, on the underlying "alignment space."

Pattern Width

How does one choose pattern width?

Choosing W too small discards available information for locating a pattern, while choosing W too large adds unnecessary noise. The Gibbs sampling algorithm, however, should be fairly robust to deviations that are not too far from the optimal W.

What is a reasonable criterion for optimal pattern width?

It can be difficult to compare multiple alignment scores directly for different choices of W, especially when all column scores are positive. One criterion for selecting W is the *Minimum Description Length Principle*.

For ungapped local multiple alignments, this is equivalent to optimizing the *BILD* score along a single high-dimensional diagonal, which can be achieved using a variation of the Smith-Waterman algorithm.

Employing the criterion of optimal BILD score, W may be modified dynamically, within a Gibbs sampling program.

Grunwald, P.D. (2007) *The Minimum Description Length Principle.* MIT Press, Cambridge, MA.

Altschul, S.F., *et al.* (2010) "The construction and use of log-odds substitution scores for multiple sequence alignment." *PLoS Comput. Biol.* **6**:e1000852.

Close Sequences

If two input sequences are too similar to one another, they can cause each other to “stick” during the sampling stage. In other words, even when they are misaligned, the current position in one sequence will cause the equivalent position in the other sequence to be selected, and vice versa.

Possible remedies

One may remove extra copies of sequences that are too similar to one another from the input set, and add them back in at a later stage. Paradoxically, this suggests that the most distantly related sequences should be aligned first.

Alternatively, one may employ a strategy analogous to the “realignment stage” in MUSCLE. The relative alignment of a set of closely related sequences can be fixed. Then segments from these sequences can be removed in tandem from the multiple alignment, and new segments (in their previously-fixed relative alignment) sampled in one pass.

Several Generalizations of the Problem

Some sequences may be missing the pattern.

Some sequences may have multiple copies of the pattern.

The sequences may contain multiple distinct patterns, either consistently ordered or in arbitrary order.

The best alignment between the consensus pattern and its occurrences within the sequences may contain gaps.