Dirichlet Mixtures, the Dirichlet Process, and the Topography of Amino Acid Multinomial Space

Stephen Altschul

National Center for Biotechnology Information
National Library of Medicine
National Institutes of Health
Bethesda, Maryland

Motivational Problem

How should one score the alignment of a single letter to a column of letters from a multiple alignment?

V
F
V
L

M

Pairwise Alignment Scores

Pairwise *substitution scores*, for the local alignment of two sequences, are implicitly of the form:

$$s_{i,j} = \log \frac{q_{i,j}}{p_i p_j}$$

where $q_{i,j}$ is the *target frequency* with which amino acids i and j correspond in biologically accurate alignments, and p_i is the *background frequency* for amino acid i in proteins.

Schwartz, R.M. & Dayhoff, M.O. (1978) In *Atlas of Protein Sequence and Structure, vol. 5, suppl. 3*, M.O. Dayhoff (ed.), pp. 353-358, Natl. Biomed. Res. Found., Washington, DC.

Karlin, S. & Altschul, S.F. (1990) *Proc. Natl. Acad. Sci. USA* **87**:2264-2268.

Generalization to Multiple Alignments

The score for aligning an amino acid to a multiple alignment column should be

$$s_i = \log \frac{q_i}{p_i}$$

where q_i is the *estimated probability* of observing amino acid i in that column.

Transformed motivational problem:
How should one estimate the twenty components of $\vec{q}$ from a multiple alignment column that may contain only a few observed amino acids?

V
F
V
L

A Bayesian Approach

Define a *prior probability distribution* over *multinomial space* for the amino acid frequency vectors that characterize real proteins.

When combined with a set of observed amino acids in a particular multiple alignment column, Bayes' Theorem implies a *posterior distribution* over multinomial space, and $\vec{q}$ may be derived by integrating over this posterior distribution.

For purely mathematical reasons, the prior distribution should be a *Dirichlet distribution*, or a *Dirichlet mixture*, because then the posterior distribution is easily calculated as another Dirichlet distribution or Dirichlet mixture.

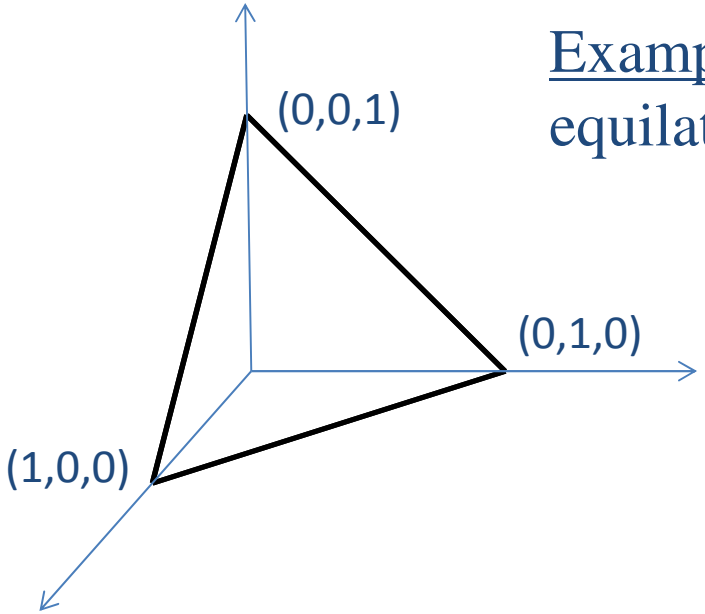
Brown, M. *et al*. (1993) In *Proc. First Int. Conf. Intelligent Systems for Molec. Biol.* L. Hunter, D. Searls, J. Shavlik, Eds., AAAI Press, Menlo Park, CA, pp. 47-55.

Multinomial Space

A *multinomial* on an alphabet of L letters is a vector $\vec{p}$ of L positive probabilities that sum to 1.

The *multinomial space* Ω_L is the space of all multinomials on L letters.

Because of the constraint on the components of a multinomial, Ω_L is $L-1$ dimensional.



Example: Ω_3 is a 2-dimensional equilateral triangle.

For proteins, we will be interested in the 19-dimensional multinomial space Ω_{20}.

The Dirichlet Distribution

Bayesian analysis will work for *any* prior, but when dealing with multinomial space, it is mathematically convenient to require the prior to be a Dirichlet distribution*.

The Dirichlet distributions are an L-parameter family of probability densities over the $(L-1)$-dimensional space Ω_L. A particular Dirichlet distribution, represented by a vector $\vec{\alpha}$ with positive components, has probability density given by:

$$\rho(\vec{x}) \;=\; Z \prod_i x_i^{\alpha_i - 1},$$

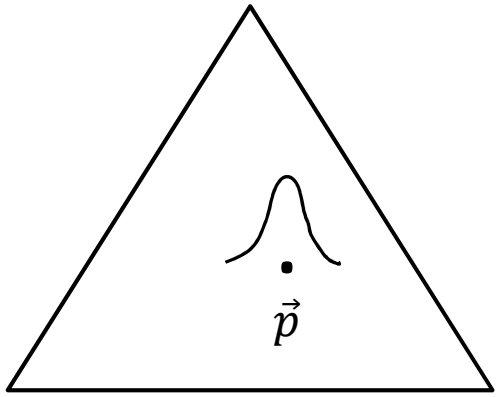
where $Z = \Gamma(\sum \alpha_i) / \prod \Gamma(\alpha_i)$ is a constant chosen so that $\rho(\vec{x})$ integrates to 1.

Note: The Dirichlet distribution with all $\alpha_i = 1$ is the uniform density.

* The *conjugate prior* for the multinomial distribution.

How to Think About Dirichlet Distributions

Define the "concentration parameter" α to be $\sum \alpha_i$. Then the center of mass of the Dirichlet distribution is $\vec{p} = \vec{\alpha}/\alpha$.



The greater α, the greater the concentration of probability near $\vec{p}$.

A Dirichlet distribution may be alternatively parameterized by: $(\vec{p}, \alpha)$.

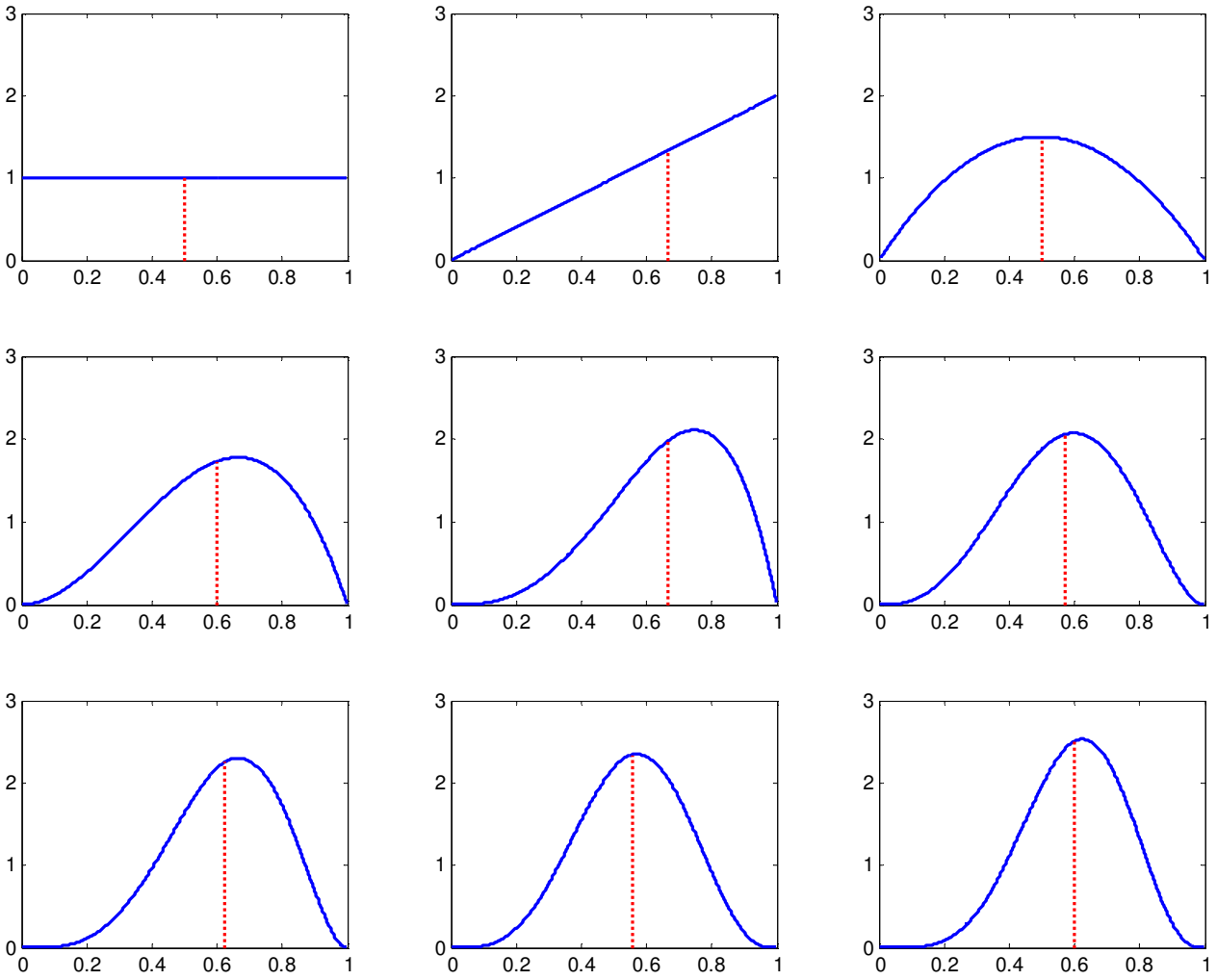
By Bayes' theorem, the observation of a single letter "a" transforms the Dirichlet prior $\vec{\alpha}$ into a Dirichlet posterior $\vec{\alpha}'$ with identical parameters, except that $\alpha'_a = \alpha_a + 1$.

Bayes at Work

Here, we begin with the uniform Dirichlet prior (1,1) for the probability of "heads", and observe its transformation, after successive observations **HTHHTHTH**, into the posteriors (2,1), (2,2), (3,2), *etc*.

At any given stage, the center of mass (i.e. the expected probability of heads) is given by:

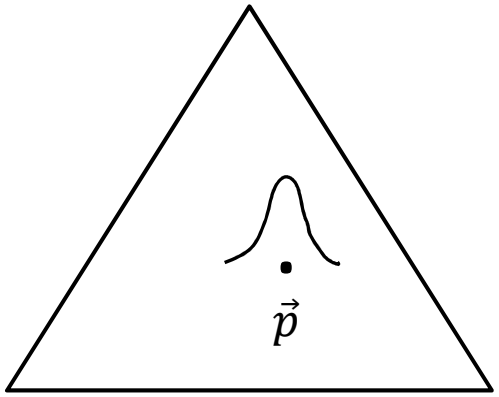
$$\frac{\#(H)+1}{[\#(H)+1]+[\#(T)+1]}$$



Note: The 2-parameter Dirichlet distributions, which take the form $Zx^{\alpha-1}(1-x)^{\beta-1}$, are also called Beta distributions.

Is the Dirichlet distribution an appropriate prior for amino acid frequencies at individual protein positions?

Although proteins as a whole have background amino acid frequencies $\vec{p}$, it is not the case that the frequencies $\vec{q}$ typical of individual protein positions tend to be clustered near $\vec{p}$.



Rather, some positions tend to be charged, some aromatic, some hydrophobic, etc., suggesting that prior probability density is concentrated in multiple regions within multinomial space.

A *Dirichlet mixture* is better able to capture this more complex prior distribution, but is still convenient for Bayesian analysis.

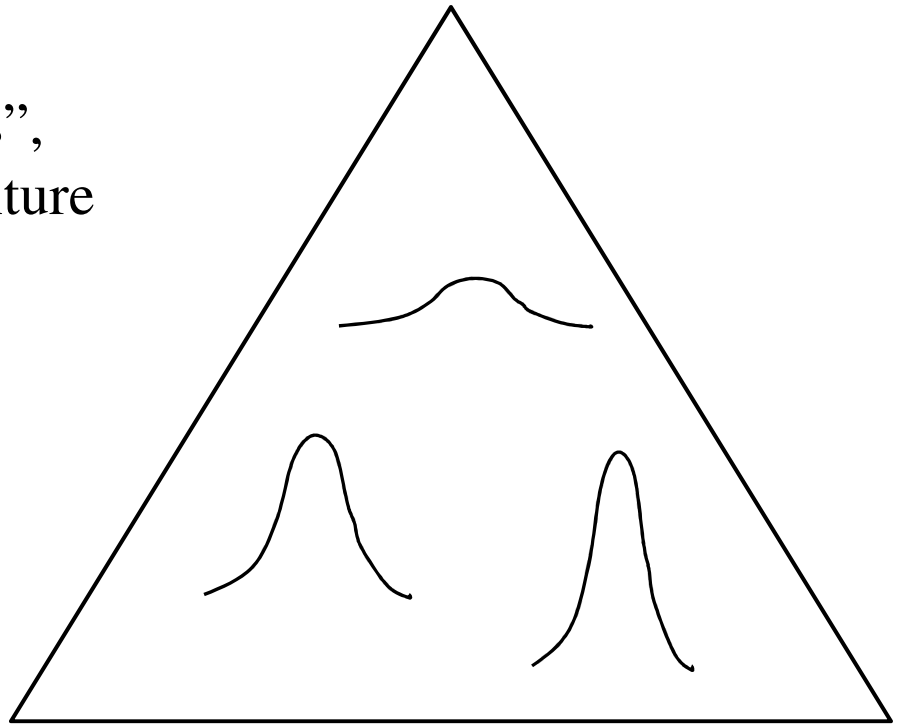
Brown, M., et al. (1993) "Using Dirichlet mixture priors to derive hidden Markov models for protein families." In: *Proc. First Int. Conf. Intelligent Systems for Mol. Biol.*, L. Hunter, D. Searls & J. Shavlik, Eds. AAAI Press, Mento Park, CA, pp. 47-55.

Dirichlet Mixtures

A *Dirichlet mixture* consists of M *Dirichlet components*, associated respectively with positive "mixture parameters" $m_1, m_2, \ldots, m_M$ that sum to 1. Only $M - 1$ of these parameters are independent.

Each Dirichlet component has the usual L free "Dirichlet parameters", so an M-component Dirichlet mixture has a total of $M(L + 1) - 1$ free parameters.

The density of a Dirichlet mixture is defined to be a linear combination of those of its constituent components.



A Dirichlet mixture max be visualized as a collection of probability hills in multinomial space.

Where do Dirichlet Mixture Priors Come From?

A Dirichlet mixture prior should capture our knowledge about amino acid frequencies within proteins. However:

No one knows how to construct a Dirichet mixture prior from first principles.

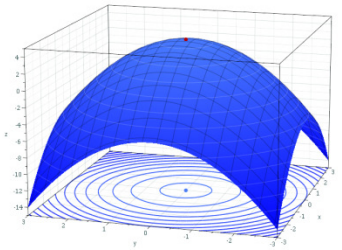
So we invert the problem: Like substitution matrices, D.M. priors may be derived from putatively accurate alignments of related sequences.

Given a large number of multiple alignment columns, we seek the *maximum-likelihood M*-component D.M., i.e. the one that best explains the data.

This is an instance of the classic, difficult problem of optimization in a rough, high-dimensional space. The only practical approaches known are heuristic.

Optimization in High-Dimensional Space

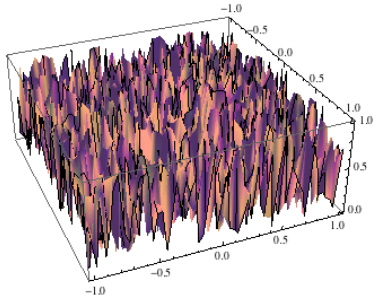
Smooth and simple landscapes



Relatively easy and fast to find optimum.

Algorithms: Newton's method; gradient descent.

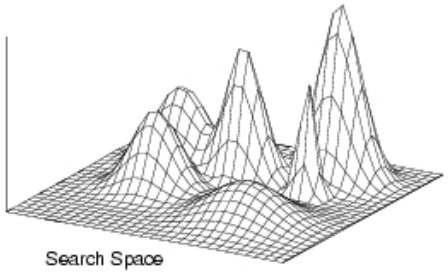
Random landscapes



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; EM; Gibbs sampling.

Success depends on details of landscape.

Difficulties: Local optima.

Images courtesy of the internet

Heuristic Algorithms

The Metropolis algorithm and simulated annealing

Metropolis, N., *et al.* (1953) "Equation of state calculations by fast computing machines." *J. Chem. Phys.* **21**:1087-1092.

Expectation maximization (EM)

Dempster, A.P., *et al.* (1977) "Maximum likelihood from Incomplete data via the EM algorithm." *J. Royal Stat. Soc., Series B* **39**:1-38.

applied to Dirichlet mixtures

Brown, M., *et al.* (1993) In: *Proc. First Int. Conf. Intelligent Systems for Molec. Biol.,* L. Hunter, D. Searls, J. Shavlik, Eds., AAAI Press, Menlo Park, CA, pp. 47-55.

Gibbs sampling

Geman, S. & Geman, D. (1984) "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images." *IEEE Trans. Pattern Analysis and Machine Intelligence* **6**:721-741.

applied to Dirichlet mixtures

Ye, X., *et al.* (2011) "On the inference of Dirichlet mixture priors for protein sequence comparison." *J. Comput. Biol.* **18**:941-954.

Nguyen, V.-A., *et al.* (2013) "Dirichlet mixtures, the Dirichlet process, and the structure of protein space." *J. Comput. Biol.* **20**:1-18.

Gibbs Sampling for Dirichlet Mixtures

Given: N multiple alignment columns

Find: The M Dirichlet components that maximize the likelihood of the data

Algorithm

1) Initialize by associating columns with components

2) Derive the parameters for each Dirichlet component from the columns assigned to it

3) In turn, sample each column into a new component, using probabilities proportional to column likelihoods

4) Iterate

How Many Dirichlet Components Should There Be?

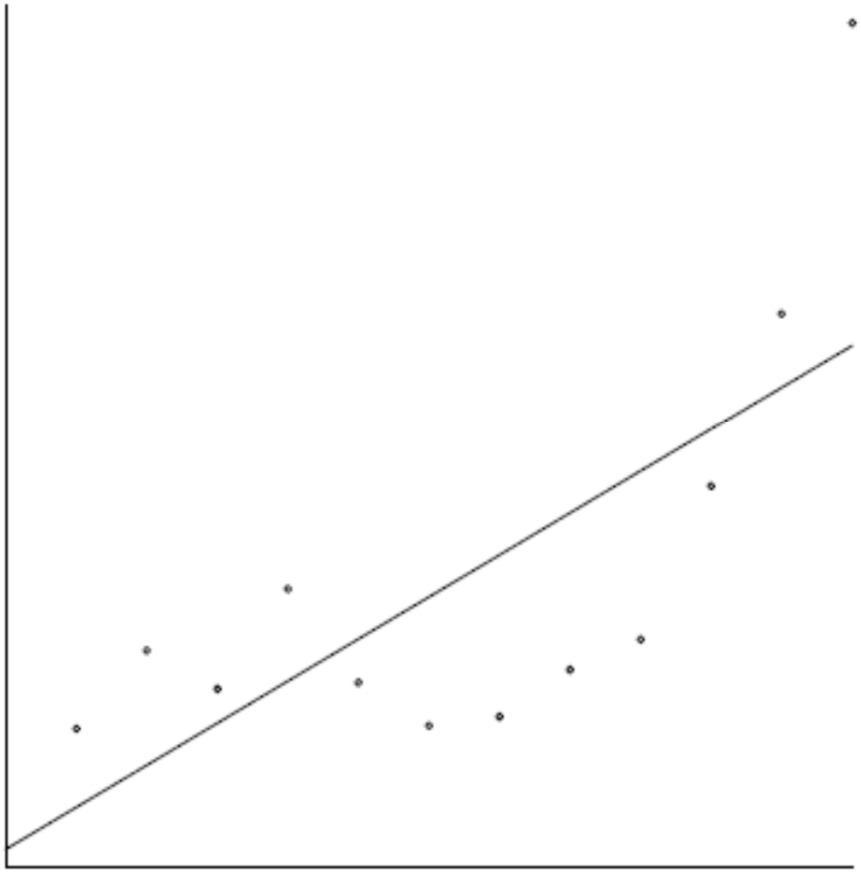
Idea: Maximize the likelihood of the data.

Problem: The more components, the greater the likelihood of the data. The criterion of maximum-likelihood *alone* leads to overfitting.

One solution: The *Minimum Description Length (MDL)* principle.

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

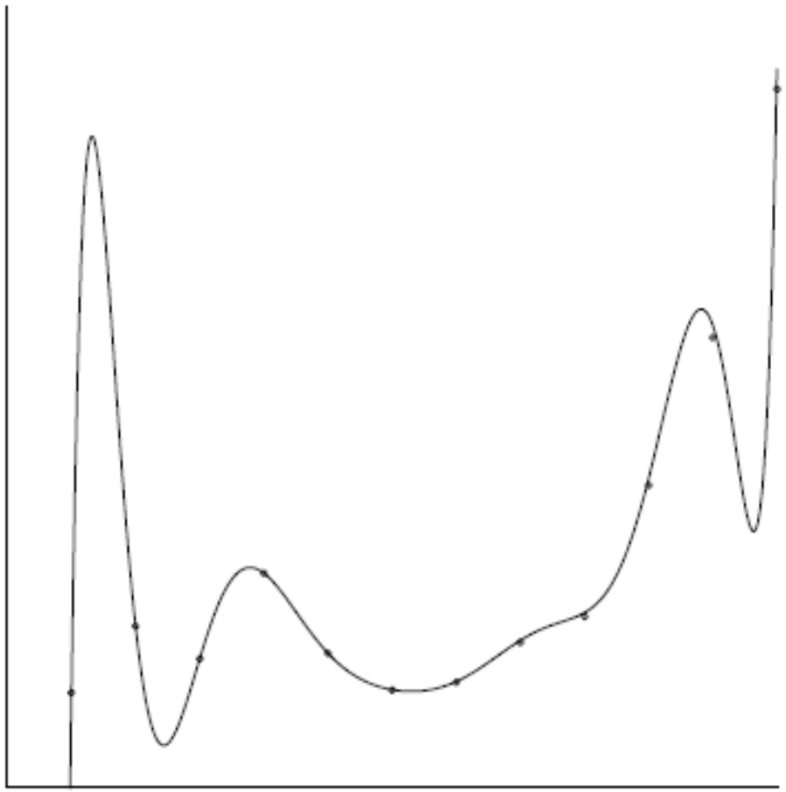
A model that is too simple underfits the data



A simple model, i.e. one with few parameters, will have low complexity but will not fit the data well.

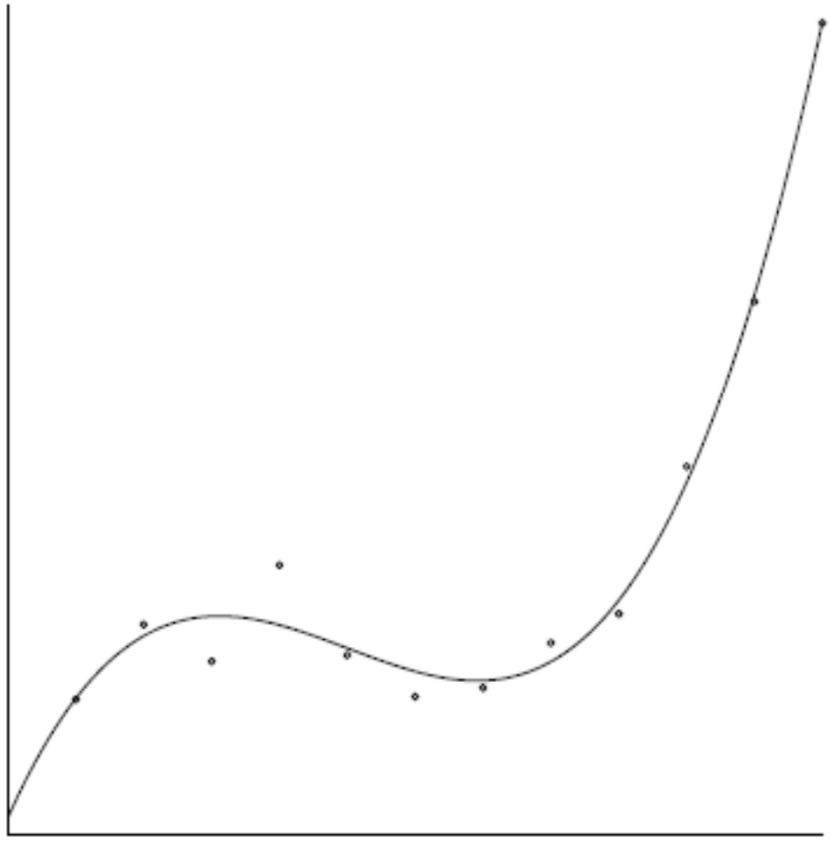
From: “A tutorial introduction to the minimum description length principle” by Peter Grünwald

A model that is too complex overfits the data



A complex model will fit the data well, but is itself long to describe.

A model with an appropriate number of parameters



Everything should be made as simple as possible, but not simpler. – Albert Einstein

A model should be as detailed as the data will support, but no more so. – MDL principle

The Minimum Description Length Principle

A set of data $\boldsymbol{S}$ may be described by a parametrized theory, chosen from a set of theories called a model, $\boldsymbol{M}$.

$\mathrm{DL}(\boldsymbol{S}|\boldsymbol{M})$, the description length of $\boldsymbol{S}$ given $\boldsymbol{M}$, is the negative log probability of $\boldsymbol{S}$ implied by the maximum-likelihood theory contained in $\boldsymbol{M}$.

MDL theory defines the *complexity* of a model, $\mathrm{COMP}(\boldsymbol{M})$. It may be thought of as the log of the effective number of independent theories the model contains.

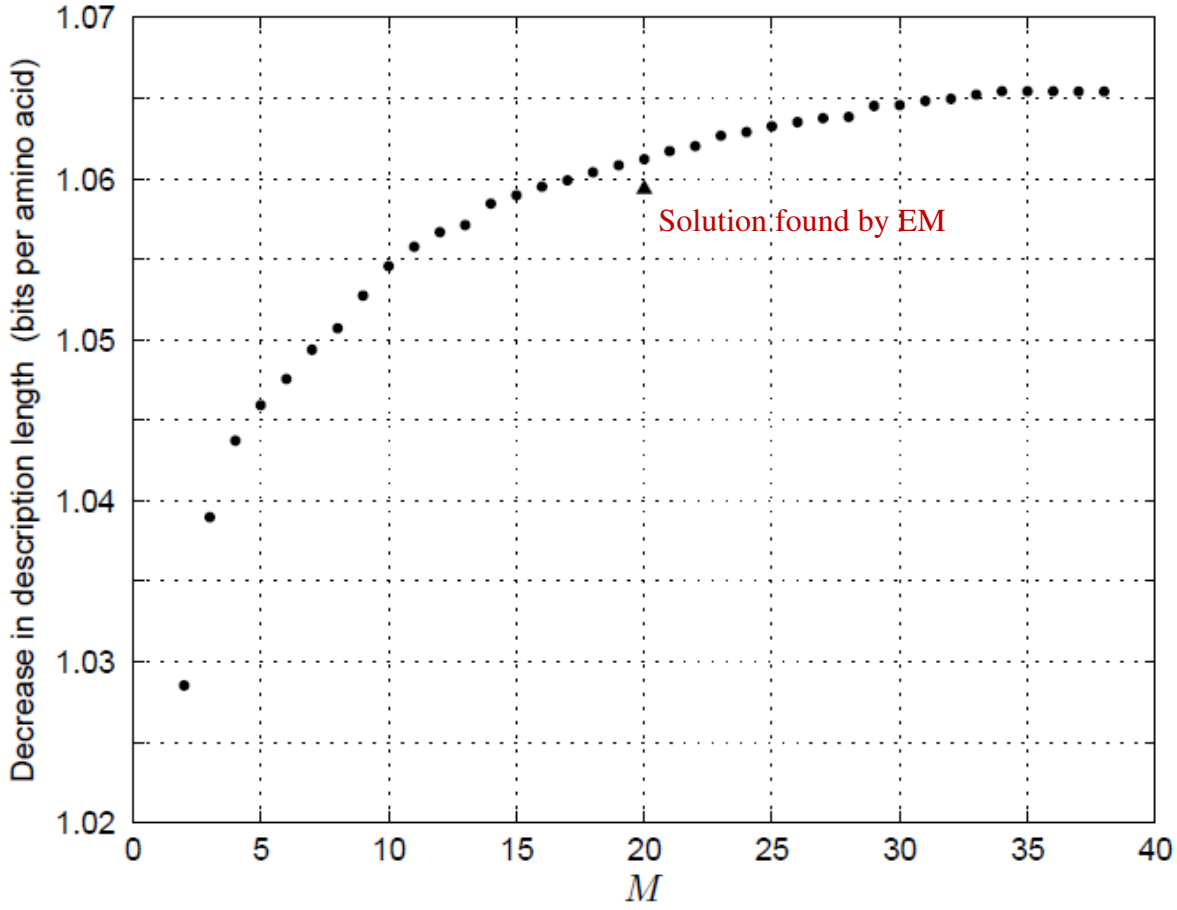
The MDL principle asserts that the best model for describing $\boldsymbol{S}$ is that which minimizes: $\mathrm{DL}(\boldsymbol{S}|\boldsymbol{M}) + \mathrm{COMP}(\boldsymbol{M})$.

Effect: *More complex models are penalized*

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

The Optimal Number of Dirichlet Components
(estimated using Gibbs sampling algorithm)

Data set: "**diverse-1216-uw**", containing 315,585 columns with an average of 76.0 amino acids per column, from: https://compbio.soe.ucsc.edu/dirichlets/index.html



Decrease in total description length:

1.0654 bits/a.a.

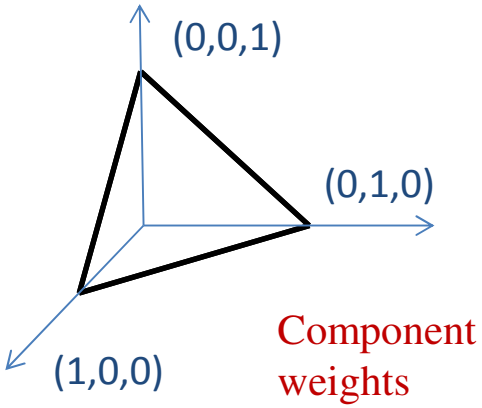
using a 35-component Dirichlet mixture

Problem: How effective is the algorithm at finding a maximum-likelihood DM?

Ye, X., *et al.* (2011) *J. Comput. Biol.* **18**:941-954.

The Dirichlet Process

Many distributions may be modeled as *mixtures* of an *underlying distribution*. For example, the distribution of points along a line may be modeled by a mixture of normal distributions.



The Dirichlet Process (DP) is used to model mixtures with an unknown or an unbounded number of components. The name derives from a generalization of the Dirichlet distribution to an infinite number of dimensions, to model the *weights* of these components.

A DP may be thought of as assigning a generalized prior probability to mixtures with an infinite of components.

A DP is completely specified by two elements:

A prior distribution H over the parameters of the underlying distribution

A positive real hyperparameter, which we will call γ, which defines a prior on the weights of the components

The smaller γ, the greater the implied concentration of weight in a few components.

Antoniak, C.E. (1974) *Ann. Stat.* **2**:1152-1174.

The Chinese Restaurant Process

A restaurant with an infinite number of tables.

People enter sequentially and sit randomly at tables, following these probabilities:

At an occupied table k, with probability proportional to the number of people N_k already seated there;

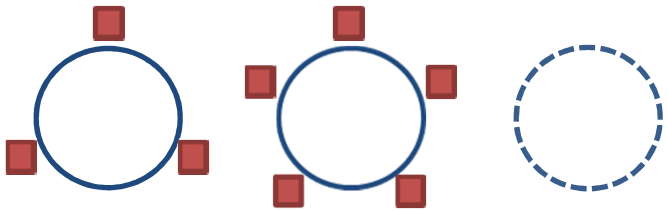
At a new, unoccupied table, with probability proportional to γ.

Example: 8 people already seated: 3 at Table 1; 5 at Table 2; $\gamma = 2$.

Probability to sit at Table 1: 0.3

Probability to sit at Table 2: 0.5

Probability to sit at a new table: 0.2



Each table corresponds to a component, with its parameters chosen randomly according to the prior distribution H.

The proportion of people seated at a table corresponds to its weight.

Ferguson, T.S. (1973) *Ann. Stat.* **1**:209-230.

Dirichlet-Process Modifications to the Gibbs Sampling Algorithm

When sampling a column C into a component:

If C was the only column associated with its old component, abolish that component.

Allow C to seed a new component, with probability proportional to γ. This may be calculated by integrating $\gamma\ \mathrm{Prob}(C|\vec{q})\mathrm{Prob}(\vec{q}|\vec{\alpha})$ over Ω_{20} and Dirichlet parameter space, using the prior density H.

If a new component is created, sample its parameters, as below.

When calculating Dirichlet parameters for a component:

Sample the parameters from the posterior distribution implied by H and the columns assigned to the component.

Component Likelihoods

Number of columns associated with component k: n_k

Number of occurrences of amino acid j in column: c_j

Total number of amino acids in column: c

Dirichlet parameters for component k: $\alpha_{k,j}$

Sum of Dirichlet parameters for column k: α_k

DP hyperparameter for prior over Dirichlet-distribution parameter space: β

DP hyperparameter for component weights: γ

Background frequency for amino acid j: p_j

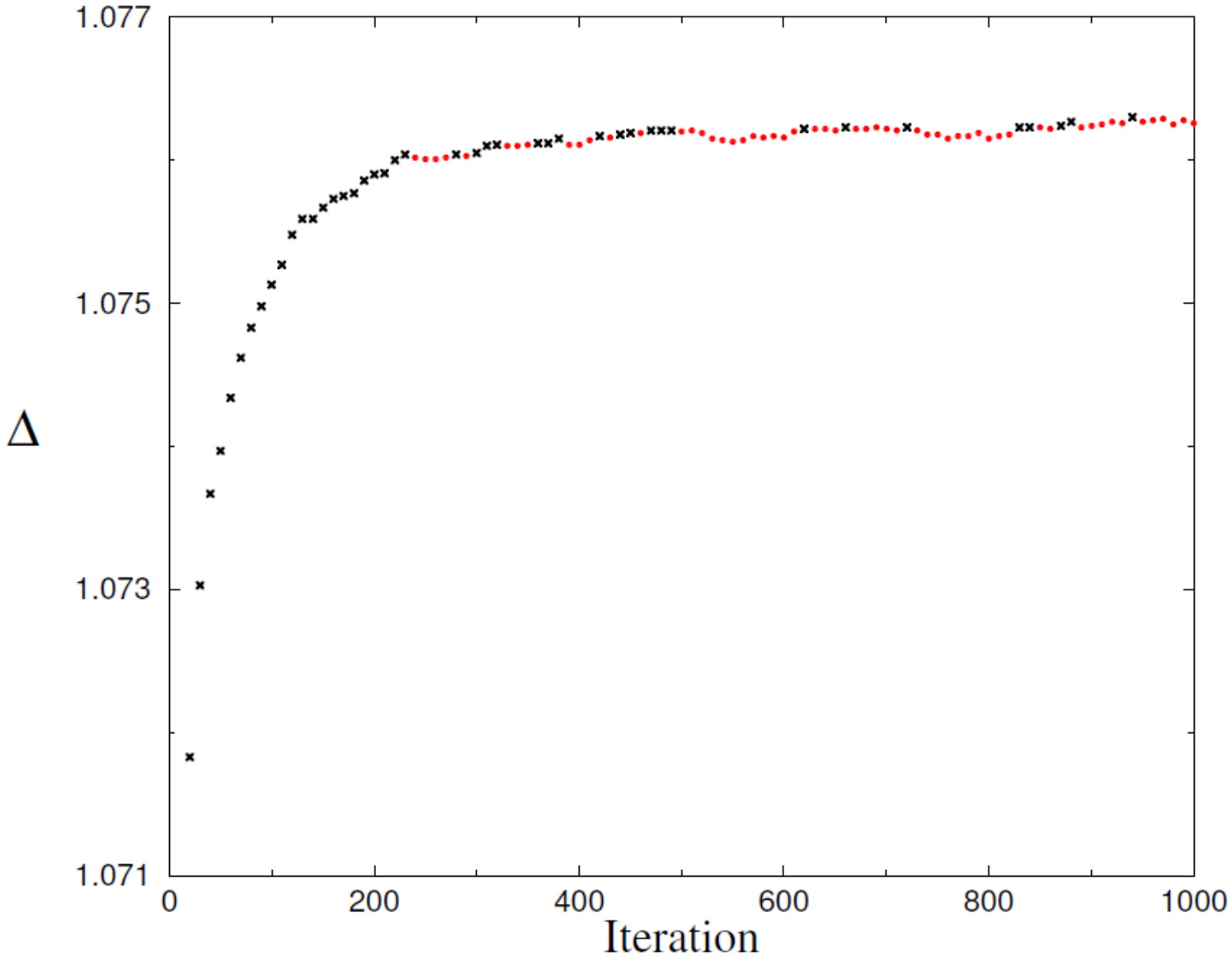
$$\text{Prob(component } k) \quad \propto \quad n_k \frac{\Gamma(\alpha_k)}{\Gamma(\alpha_k + c)} \prod_{j=1}^{20} \frac{\Gamma(\alpha_{k,j} + c_j)}{\Gamma(\alpha_{k,j})}$$

$$\text{Prob(new component)} \quad \propto \quad \gamma \frac{\Gamma(\beta)}{\Gamma(\beta + c)} \prod_{j=1}^{20} \frac{\Gamma(\beta p_j + c_j)}{\Gamma(\beta p_j)}$$

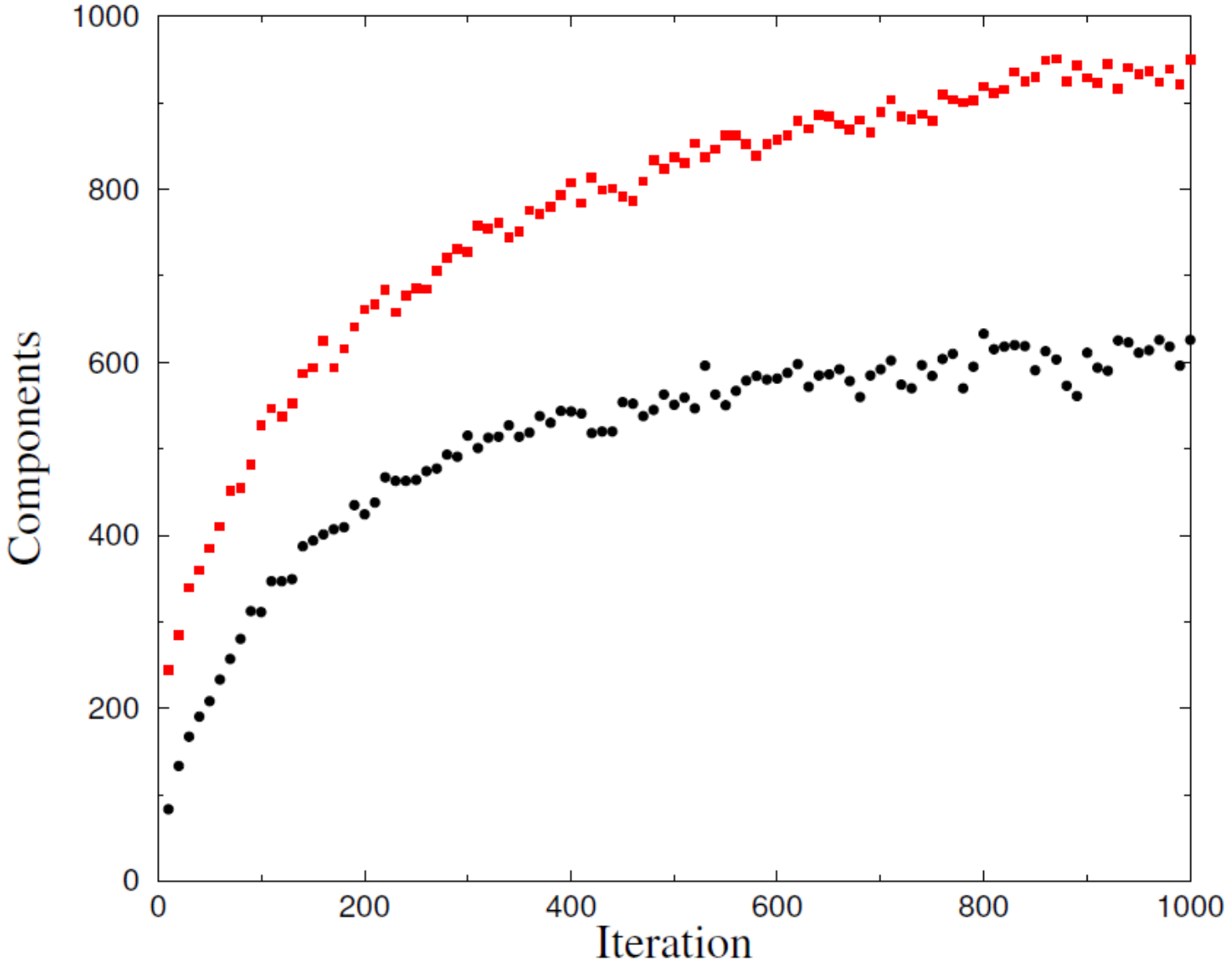
Decrease in Total Description Length as a Function of the Dirichlet Process Hyperparameters β and γ

β	γ	Best Δ (bits/a.a.)	Number of components	Iteration found	β	γ	Best Δ (bits/a.a.)	Number of components	Iteration found
100	5	1.0756	470	950	600	5	1.0757	361	980
	10	1.0758	520	860		10	1.0760	406	960
	20	1.0759	644	980		20	1.0762	471	930
	40	1.0760	689	980		40	1.0762	481	980
	60	1.0760	720	800		60	1.0763	533	910
	80	1.0760	717	630		80	1.0762	526	900
	100	1.0760	808	1000		100	1.0762	541	780
200	5	1.0757	449	1000	800	5	1.0757	341	1000
	10	1.0759	498	840		10	1.0759	378	1000
	20	1.0761	586	960		20	1.0761	431	960
	40	1.0761	597	600		40	1.0761	466	830
	60	1.0762	665	750		60	1.0762	472	830
	80	1.0762	709	870		80	1.0762	471	730
	100	1.0762	679	590		100	1.0762	499	760
400	5	1.0757	400	980	1000	5	1.0755	314	1000
	10	1.0760	452	910		10	1.0758	350	1000
	20	1.0762	505	960		20	1.0759	375	990
	40	1.0763	562	860		40	1.0760	429	860
	60	1.0763	588	1000		60	1.0761	433	910
	80	1.0763	603	990		80	1.0761	447	910
	100	1.0763	623	940		100	1.0761	444	990

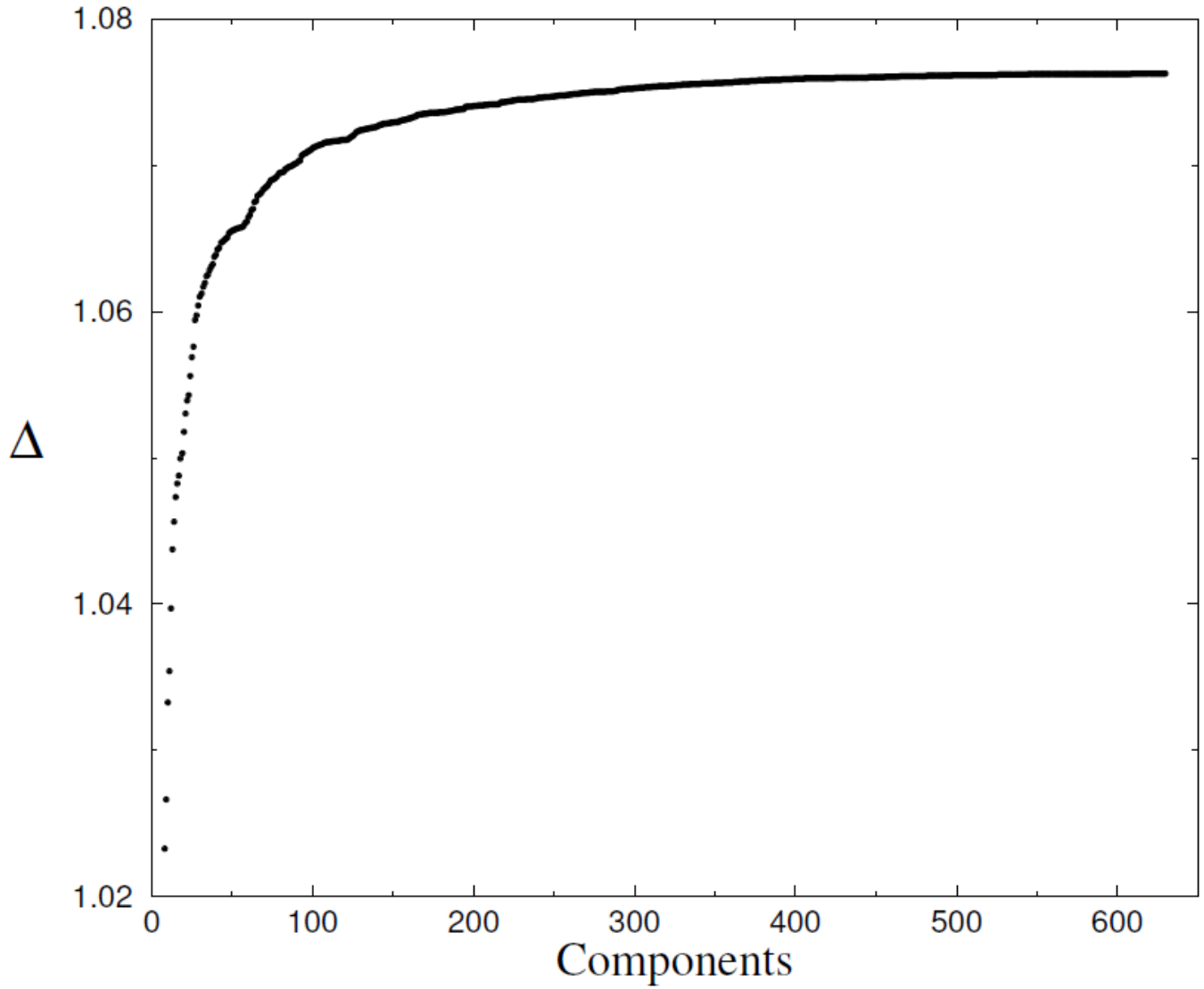
Decrease in Total Description Length as a Function of DP-Sampler Iteration ($\beta = 400;\ \gamma = 100$)



Total Number of Components, and Number Supported by the MDL Principle, as a Function of DP-Sampler Iteration



Tradeoff Between Number of Dirichlet Components and Decrease in Total Description Length per Amino Acid



Visualizing Dirichlet Mixture Components

Reorder the amino acids: RKQEDNHWYFMLIVCTSAGP

Represent the target frequency q_j for an amino acid by a symbol for its implied log-odds score $s_j = \log_2(q_j/p_j)$ as follows:

$s_j > 2$	$\sigma_j =$ The amino acid's one-letter code, in upper case
$2 \geq s_j > 1$	$\sigma_j =$ The amino acid's one-letter code, in lower case
$1 \geq s_j > 0.5$	$\sigma_j =$ "+"
$0.5 \geq s_j > -1$	$\sigma_j =$ " "
$-1 \geq s_j > -2$	$\sigma_j =$ "."
$-2 \geq s_j > -4$	$\sigma_j =$ "–"
$-4 \geq s_j$	$\sigma_j =$ "="

A Reordered Subset of a 134-Component Dirichlet Mixture

Rank	w (%)	α_k	RKQEDNHWYFMLIVCTSAGP
69	0.51	30.7	R .-=- ----------
23	1.20	26.7	R+ .. . .
124	0.26	35.3	K.--..---.---------
15	1.49	27.0	rK+ . -.- .-.- -.
3	2.82	27.0	rk+ - + -.
89	0.41	0.4	RKq - +- -=-====--
24	1.16	33.0	+++ -.-..-.. +a .
7	1.91	62.7	rkq+ ..- .... .-
2	3.18	59.5	++++ .
91	0.41	164.5	+kqe+ .-------- a..
6	1.95	106.3	+kqe+ ..-..--- ..
18	1.37	37.2	+kqE+ --=----- ..
25	1.13	36.1	+k+ +n -------- +
19	1.33	97.6	+++++ ... . +
41	0.80	74.4	++edn -.------
60	0.61	22.7	Q+ ... .
83	0.45	6.9	. qE+ .-- ---.
51	0.67	57.6	. qEd -------- .-
5	2.15	34.3	+E . .

The Topography of Amino Acid Multinomial Space

Rank	w (%)	α_k	RKQEDNHWYFMLIVCTSAGP	Rank	w (%)	α_k	RKQEDNHWYFMLIVCTSAGP
69	0.51	30.7	R .-=- ----------	93	0.40	24.9	-=-===- FmL . ---=-
23	1.20	26.7	R+	65	0.55	52.7	==-===- fmL+ .-.--
124	0.26	35.3	K.--..---.---------	92	0.40	34.0	==-==-. fmliv+ . ..
15	1.49	27.0	rK+ . -.- .-.- -.	4	2.56	37.2	....-.. +mL+ .. -.
3	2.82	27.0	rk+ - + -.	64	0.57	32.6	========--+mLI --=-==
89	0.41	0.4	RKq - +- -=-=====--	11	1.67	49.0	.-.---. ++liv . -.
24	1.16	33.0	+++ -.-..-.. +a .	125	0.25	6.6	-- -----. M+ .------
7	1.91	62.7	rkq+ ..- .... .-	43	0.76	14.8	.- -=-. Ml+ + . --
2	3.18	59.5	++++ .	39	0.82	6.6	-=-======= mL+ .-=-=-
91	0.41	164.5	+kqe+ .-------- a..	16	1.44	67.9	++++ .
6	1.95	106.3	+kqe+ ..-..--- ..	76	0.48	22.2	========-- mliv - =-
18	1.37	37.2	+kqE+ --=----- ..	105	0.32	28.0	==-===-.- mli++ .a.=
25	1.13	36.1	+k+ +n -------- +	35	0.97	61.8	==-=====-- +L .---==
19	1.33	97.6	+++++ ... . +	54	0.65	82.9	---===--.. +lIv --.=-
41	0.80	74.4	++edn -.------	99	0.37	47.9	========-- lIV.-=.=-
60	0.61	22.7	Q+	29	1.00	22.3	========- +Iv.-=-=-
83	0.45	6.9	. qE+ .-- ---.	106	0.32	3.5	-=====-===-. IV--=---
51	0.67	57.6	. qEd --------- .-	72	0.49	54.4	========-- IV -=-==
5	2.15	34.3	+E . .	42	0.78	52.4	-=-===-.. IV .-.--
85	0.44	43.2	-- E .---------------	8	1.86	10.4	.-... iv . -.
95	0.39	63.2	+e+ -------- s -	9	1.85	37.2	.. iv ..
27	1.04	107.4	+Ed --------- .	71	0.50	70.9	-=-=====. iV .- --
101	0.35	0.4	=- ED =-=-=== =-= =	46	0.71	17.4	========---. iV - =-
86	0.44	43.3	eD --=---=--	61	0.59	36.3	==-=====. iV+ .a-.
129	0.21	23.0	-- eD --------.----	22	1.22	23.4	.-.-- . + v+T+ ..
10	1.68	38.4	+Dn	31	0.99	4.7	-=.-=... m +C a .
126	0.24	13.2	--. D - -.-.. ++ .	34	0.97	34.7	----=-- ++ +c a -
79	0.47	61.8	Dn -----=-. +. .	68	0.52	34.9	==-====--- . +c A.-
117	0.29	24.9	-.. DN -------.. -.-	32	0.98	34.9	.-.--. ... .. + A -
48	0.68	26.8	dN+-.-----. .	74	0.48	9.7	==-==.-=-. . vCTsa..
109	0.32	25.3	---- N =-=--=-.-.---	73	0.48	38.1	.-... .... .. c+sa .
98	0.37	29.9	.. -.N+. .....+ .--	131	0.19	22.4	-----.-=-----.c+Sa -
17	1.38	27.8	+++ nh y .-. .	103	0.34	5.2	-=---.=.---==-c sA+.
63	0.58	70.7	++++ . +	90	0.41	0.4	- -. --=-==-C+s g+
70	0.51	21.5	. .. H y	21	1.28	13.6	.. . ++ + ...++s
58	0.62	4.7	hWYf --.. .	102	0.35	13.1	-=-=----=----. T+---
96	0.38	1.4	-=-==-+WYF---= ===-=	47	0.69	27.3	.-.-- .---.... Ts ..
13	1.63	23.8	...--.+wYF ...--	97	0.38	35.6	. . +n.--------.Ts. .
118	0.29	27.9	-=----.W+ ..------=-	44	0.75	2.7	- nh-= -=- ts +
77	0.47	26.6	Wy+ ..	12	1.67	44.1	++ . . . .ts.-.
130	0.19	38.5	------ WyF . ...--	28	1.03	49.4	n . ..... +s
114	0.30	24.8	-=-==- wYF ...---=-	94	0.39	20.3	----- -------- +S..-
1	3.44	29.6	. wyf++ .	75	0.48	23.7	. ------- +S
128	0.21	21.0	.-----.W+fm++ .-.--	116	0.29	11.0	--.. -----===. s G
80	0.47	32.6	-=-==- +Y+----.----=	120	0.28	46.1	--.-- ..--.--. saG.
38	0.84	24.6	-=-==--+yF ...-----	132	0.18	39.3	----------.-- +. AG-
81	0.46	11.3	-=-==- +Yf++iv --.--	112	0.31	24.2	======--==--==- - aG-
123	0.27	11.7	. + y+m .	121	0.27	90.2	----------=-.- G-
53	0.66	33.1	==-===- +F++i+ .-.=-	115	0.29	14.6	--.--..------. a P

Group A:

The main ridge

Another Section of the Main Ridge

85	0.44	43.2	-- E .--------------
95	0.39	63.2	+e+ -------- s -
27	1.04	107.4	+Ed -------- .
101	0.35	0.4	=- ED =-=-=== ==- =
86	0.44	43.3	eD --=--=--
129	0.21	23.0	-- eD --------.----
10	1.68	38.4	+Dn . .....
126	0.24	13.2	--. D - -.-.. ++ .
79	0.47	61.8	Dn -----=-. +. .
117	0.29	24.9	-.. DN -------.. -.-
48	0.68	26.8	dN+-.-----. .
109	0.32	25.3	---- N =-=--=-.-.---
98	0.37	29.9	.. -.N+. .....+ .--
17	1.38	27.8	+++ nh y .-. .
63	0.58	70.7	++++ . +
70	0.51	21.5	. .. H y ... ...
58	0.62	4.7	hWYf --.. .
96	0.38	1.4	-=-==-+WYF---= ===-=
13	1.63	23.8	...--.+wYF ...--

Group B: Hydrophylic Positions Favoring Glycine or Proline

Rank	w (%)	α	RKQEDNHWYFMLIVCTSAGP
100	0.36	32.5	+k+ n -----=--. .G.
82	0.46	38.1	. dn -.---=--. .G.
78	0.47	100.0	. n ----===-- -G.
55	0.63	83.2	++ -------- G
30	1.00	50.3	+ ...-. G
57	0.62	82.6	.------.. .G
113	0.31	43.1	-------- gP
45	0.72	75.9	+d+ ..-----. + +p
108	0.32	31.7	. d+ -------. s P
127	0.21	77.4	d+ -------.ts. p
56	0.63	69.9	ed -------- P
110	0.31	84.8	+k+e+ -------- p
119	0.28	9.2	rk d+ --= -=-- . p
50	0.67	41.6	rk+ -.-..... p
33	0.98	85.6	+ . . . . p
59	0.62	66.7	+ + -------- .P
87	0.44	48.5	. .--------. ..P

Group C: Positions Favoring Single Amino Acids

Rank	w (%)	α	RKQEDNHWYFMLIVCTSAGP
111	0.31	16.3	Q -. --- .---..---
67	0.52	52.8	.. D --------. ...
88	0.41	60.4	==-.D.-==========-=-
122	0.27	34.9	--.--.H-.---=-----=-
133	0.18	24.5	.. . .C ...
134	0.16	59.0	==-=----------C-----
14	1.60	41.8	. a .
66	0.55	40.3	========-----. . A -
26	1.06	43.8	g
62	0.59	27.9	.-.--.. ... . G.
49	0.68	112.4	-..-. .=-=-===--.-G-
36	0.94	80.3	=====-========-=--G=
20	1.32	66.2	p
40	0.82	44.8	.. .P
37	0.93	42.9	.... P
107	0.32	17.3	------. . -P
84	0.44	62.5	....-..... .. .P
52	0.66	51.7	----------------.-P
104	0.34	0.0	. --H-. =-==C-.-Gp

Slice Sampling γ

β	Mean and standard deviation of γ	Best Δ (bits/a.a.)	Number of components	Iteration found	γ
100	199.7 ± 13.0	1.0760	767	280	200
200	183.9 ± 6.8	1.0762	721	580	166
400	129.6 ± 4.3	1.0763	608	790	128
600	95.5 ± 3.9	1.0762	537	930	94
800	82.2 ± 3.5	1.0762	482	990	83
1000	66.8 ± 2.8	1.0760	442	940	69

Collaborators

National Center for Biotechnology Information

Xugang Ye
Yi-Kuo Yu

University of Maryland, College Park

Viet-An Nguyen
Jordan Boyd-Graber