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

Why Multiple Alignment?

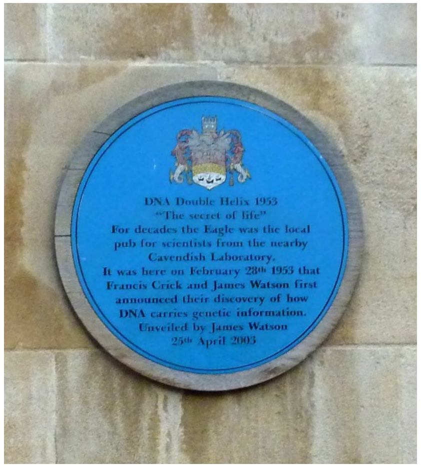
Why Multiple Alignment?

British and American bombers, WWII

The Eagle Pub, Cambridge

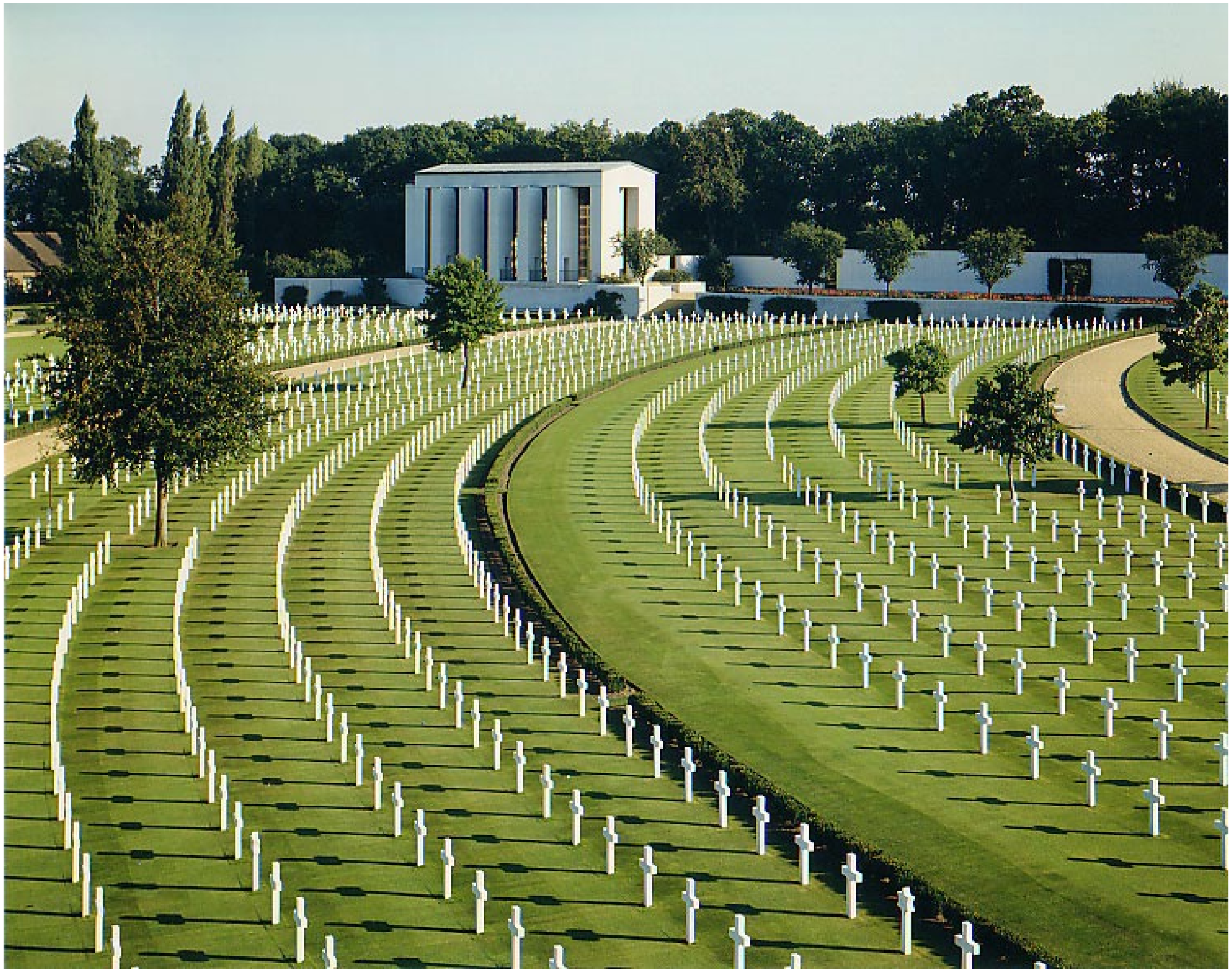
Graffiti on ceiling, written by members of the RAF and the US 8th Airforce

The Eagle Pub, Cambridge



DNA Double Helix 1953
"The secret of life"
For decades the Eagle was the local
pub for scientists from the nearby
Cavendish Laboratory.
It was here on February 28th 1953 that
Francis Crick and James Watson first
announced their discovery of how
DNA carries genetic information.
Unveiled by James Watson
25th April 2003

Graffiti on ceiling, written by members of the RAF and the US 8th Airforce



American military cemetery, Cambridge, England

A portion of a multiple alignment

```
...GEALGRLLVVYPWTQ...
...KECFTKFLSAHHDIA...
...VVFYTSILEKAPAAK...
...VDILVKFLTGTPAAQ...
...AEGLERTLHSFPTTK...
```

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 Substitution Scores

A	4																			
R	-1	5																		
N	-2	0	6																	
D	-2	-2	1	6																
C	0	-3	-3	-3	9															
Q	-1	1	0	0	-3	5														
E	-1	0	0	2	-4	2	5													
G	0	-2	0	-1	-3	-2	-2	6												
H	-2	0	1	-1	-3	0	0	-2	8											
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4										
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4									
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5								
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5							
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6						
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7					
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4				
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5			
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11		
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4
	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V

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

Log-odds scores

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.

Henikoff, S. & Henikoff, J.G. (1992) *Proc. Natl. Acad. Sci. USA* **89**:10915-10919.

Generalization of Log-Odds Scores

Score for aligning amino acid *i* to a multiple alignment column:

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

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

Generalization of Log-Odds Scores

Score for aligning amino acid *i* to a multiple alignment column:

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

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

Transformed motivational problem:
How should we estimate $\vec{q}$ from a column that may contain only a few observed amino acids?

V
F
V
L

Generalization of Log-Odds Scores

Score for aligning amino acid *i* to a multiple alignment column:

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

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

Transformed motivational problem:
How should we estimate $\vec{q}$ from a column that may contain only a few observed amino acids?

V
F
V
L

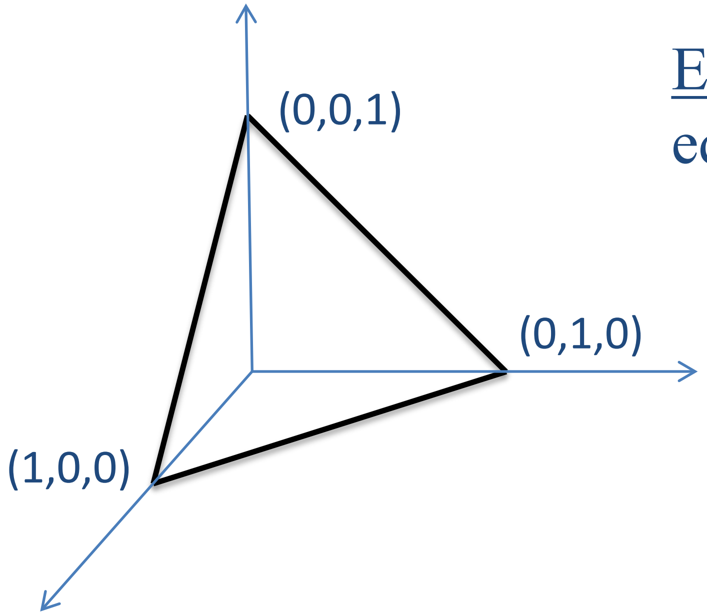
Enter Bayes…

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.

Ω_L is $L - 1$ dimensional because of the constraints on $\vec{p}$.



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

An L-parameter family of probability densities over the $(L-1)$-dimensional space Ω_L.

The Dirichlet distribution with positive parameters $\vec{\alpha}$ has density:

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

where Z is a constant chosen so that $\rho(\vec{x})$ integrates to 1.

The Dirichlet Distribution

An L-parameter family of probability densities over the $(L-1)$-dimensional space Ω_L.

The Dirichlet distribution with positive parameters $\vec{\alpha}$ has density:

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

where Z is a constant chosen so that $\rho(\vec{x})$ integrates to 1.

Johann Peter Gustav Lejeune Dirichlet

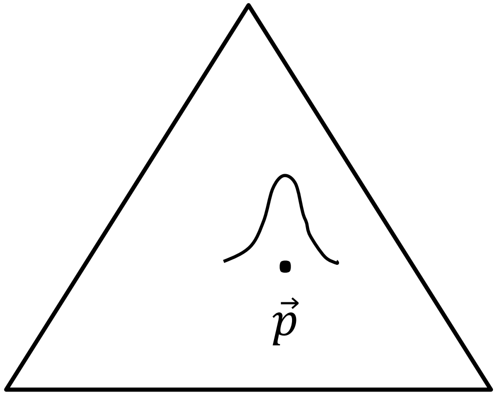
1805-1859

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

The Dirichlet distribution is the *conjugate prior* for the multinomial distribution.

How to Think About Dirichlet Distributions

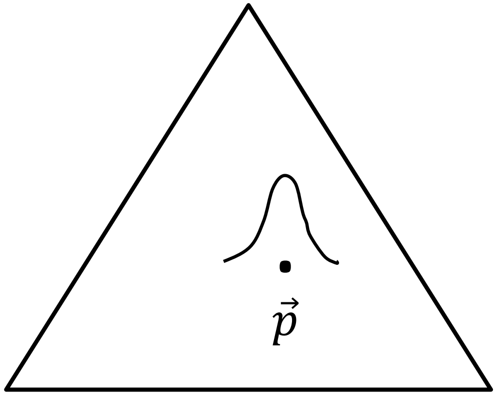
Let $\alpha = \sum \alpha_i$. The distribution's center of mass is $\vec{p} = \vec{\alpha}/\alpha$, and a greater α implies a greater concentration of mass near $\vec{p}$.



Alternative parameters: $(\vec{p}, \alpha)$.

How to Think About Dirichlet Distributions

Let $\alpha = \sum \alpha_i$. The distribution's center of mass is $\vec{p} = \vec{\alpha}/\alpha$, and a greater α implies a greater concentration of mass near $\vec{p}$.



Alternative parameters: $(\vec{p}, \alpha)$.

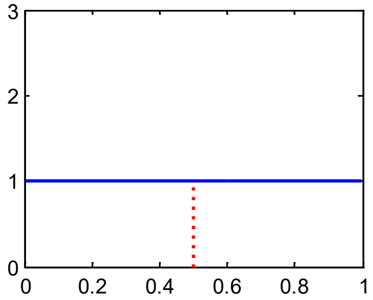
From Bayes' theorem: Observing the letter "x" transforms the Dirichlet prior $\vec{\alpha}$ into the identical posterior $\vec{\alpha}'$, except with $\alpha'_x = \alpha_x + 1$.

Thomas Bayes

1701-1761

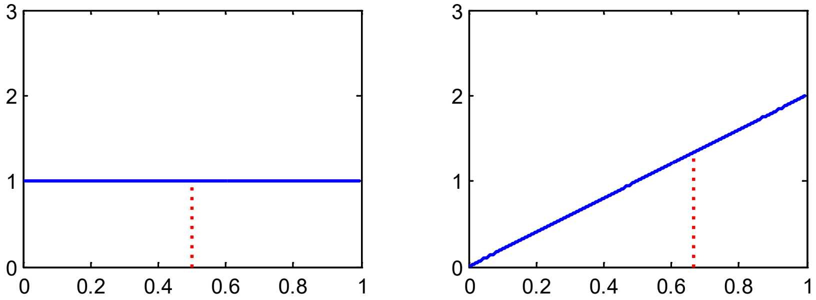
Bayes at Work

Here, we begin with the uniform Dirichlet prior (1,1) for the probability of “heads”.



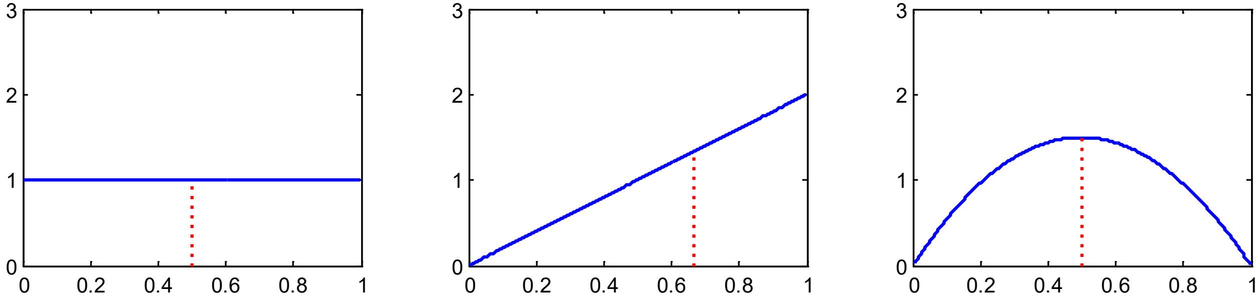
Bayes at Work

Here, we begin with the uniform Dirichlet prior (1,1) for the probability of “heads”, and observe its transformation, after the observation **H**, into the posterior (2,1).



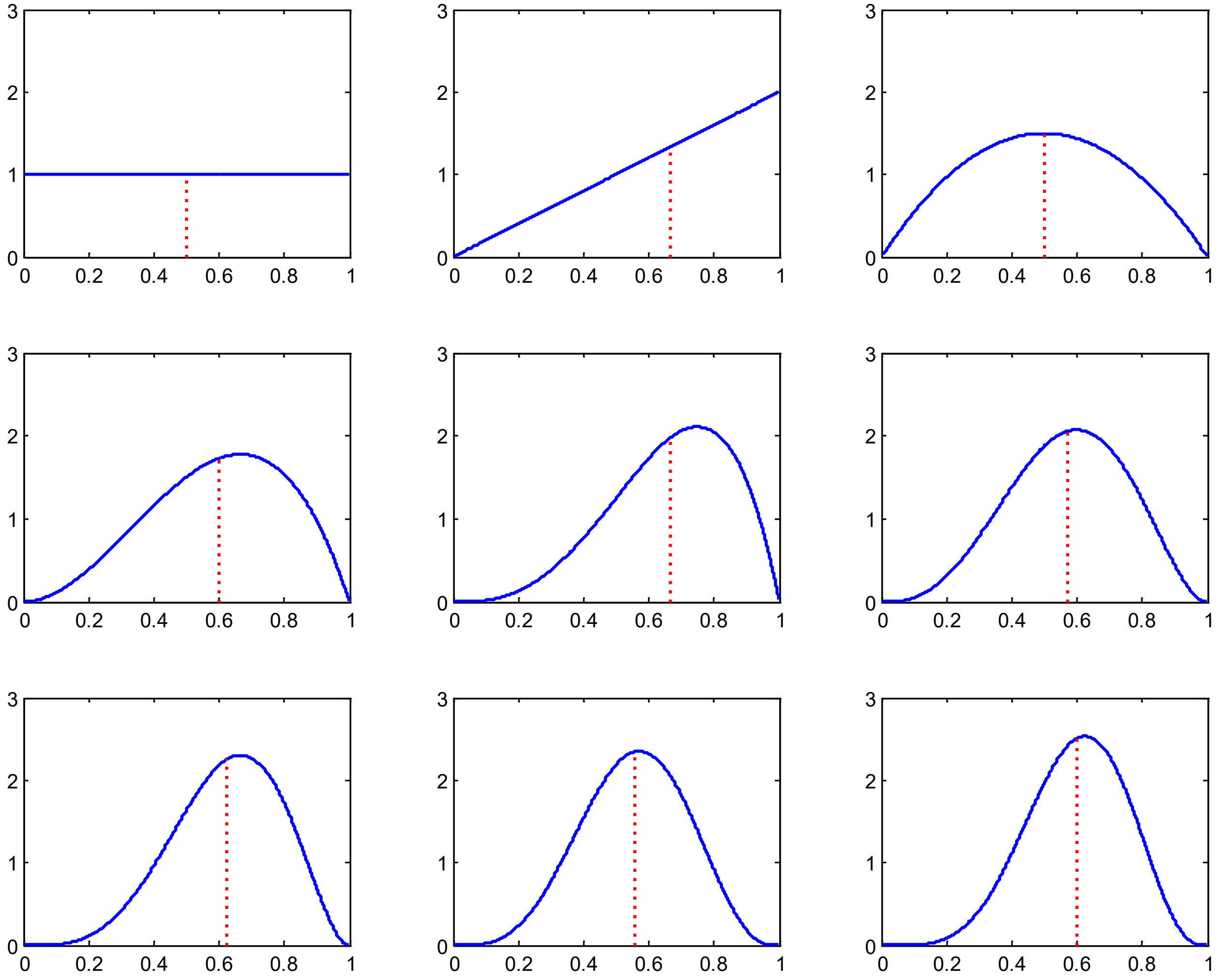
Bayes at Work

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



Bayes at Work

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

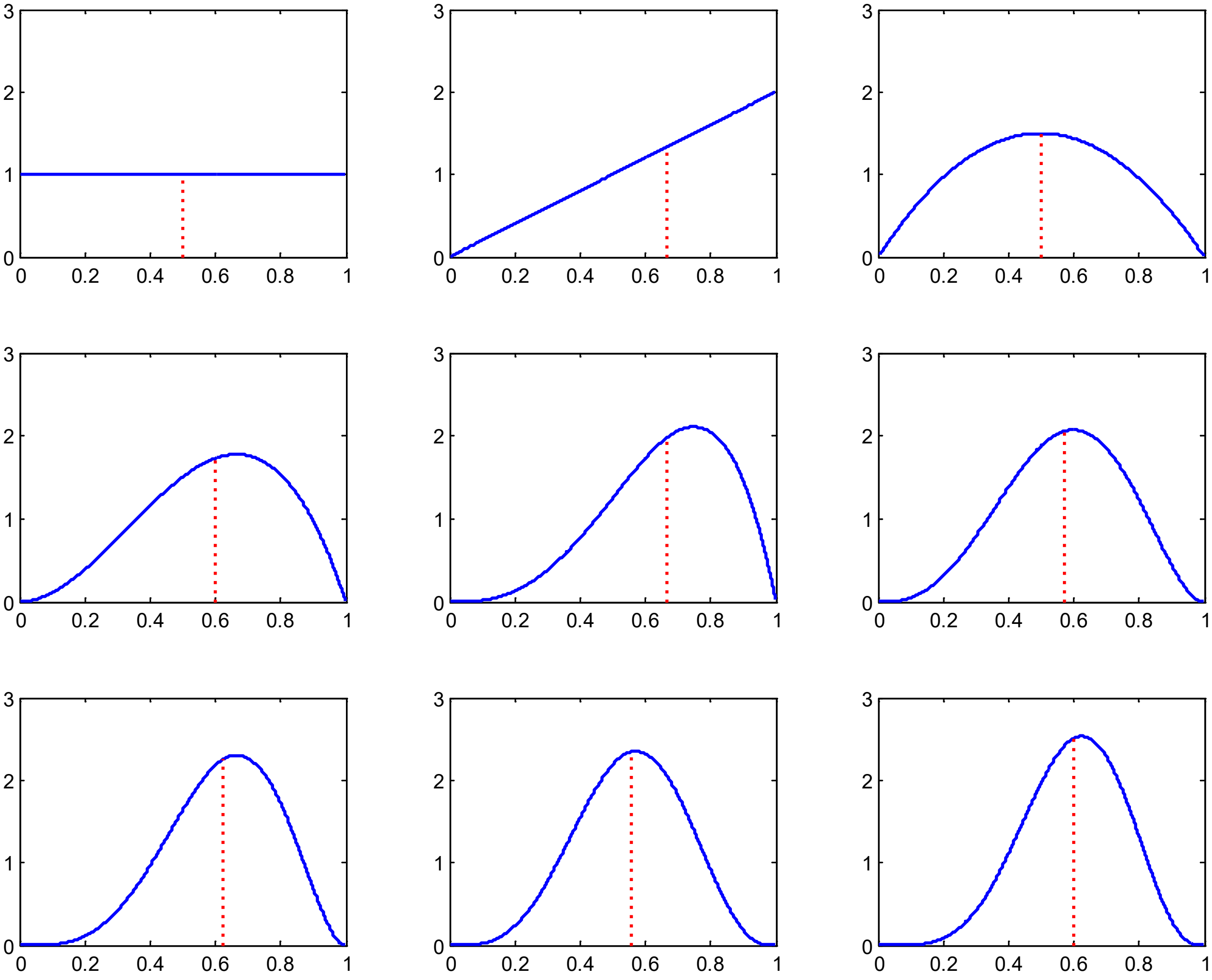


Bayes at Work

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:

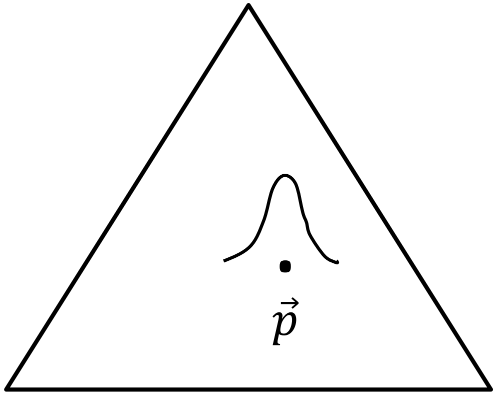
$$\frac{\#(H)+1}{[\#(H)+1]+[\#(T)+1]}$$



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

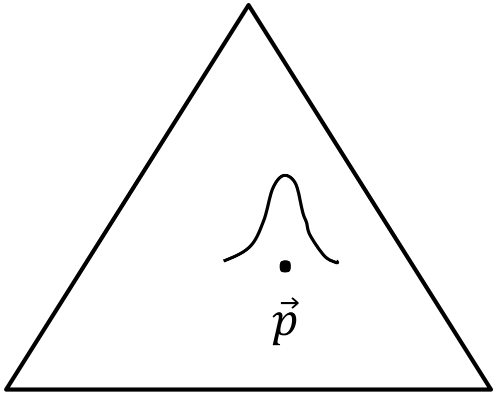
Is the Dirichlet distribution appropriate for proteins?

This distribution does not capture well our prior knowledge concerning proteins.



Is the Dirichlet distribution appropriate for proteins?

This distribution does not capture well our prior knowledge concerning proteins.



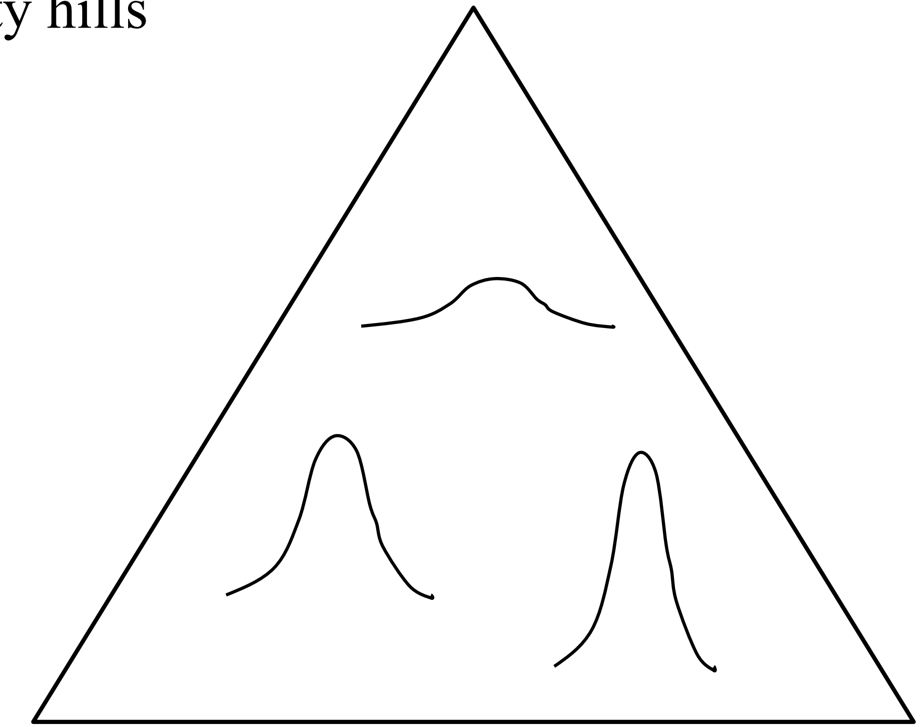
Enter the *Dirichlet mixture*…

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

The superposition of M *Dirichlet components*, with positive weights w_i that sum to 1, yielding a total of $M(L+1)-1$ free parameters.

We may visualize a Dirichlet mixture (DM) as a collection of probability hills in multinomial space.



Multiple Alignment Substitution Scores

Log-odds scores $$S(\vec{x}) = \log \frac{Q(\vec{x})}{P(\vec{x})}$$

"Bayesian Integral Log-odds" or "BILD" scores

The construction of column scores from Dirichlet mixture priors

$$Q(\vec{x}) = \sum_{i=1}^{M} w_i \frac{\Gamma(\alpha_i)}{\Gamma(\alpha_i + c)} \prod_j \frac{\Gamma(\alpha_{i,j} + c_j)}{\Gamma(\alpha_{i,j})} \qquad P(\vec{x}) = \prod_k p_{x_k}$$

where $\vec{c}$ is the amino acid count vector implied by $\vec{x}$

Assuming uniform Dirichlet priors, $S(\text{"AAACC"}) = \log(1.83) = \quad 0.87$ bits

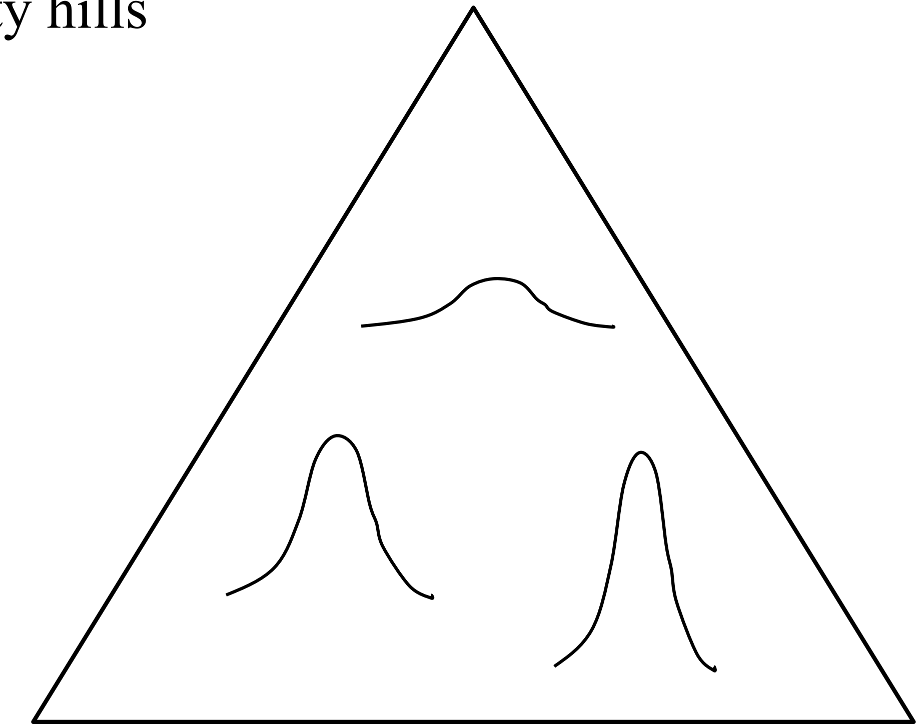
$S(\text{"AAACT"}) = \log(0.91) = -0.13$ bits

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

Dirichlet Mixtures

The superposition of M *Dirichlet components*, with positive weights w_i that sum to 1, yielding a total of $M(L+1)-1$ free parameters.

We may visualize a Dirichlet mixture (DM) as a collection of probability hills in multinomial space.

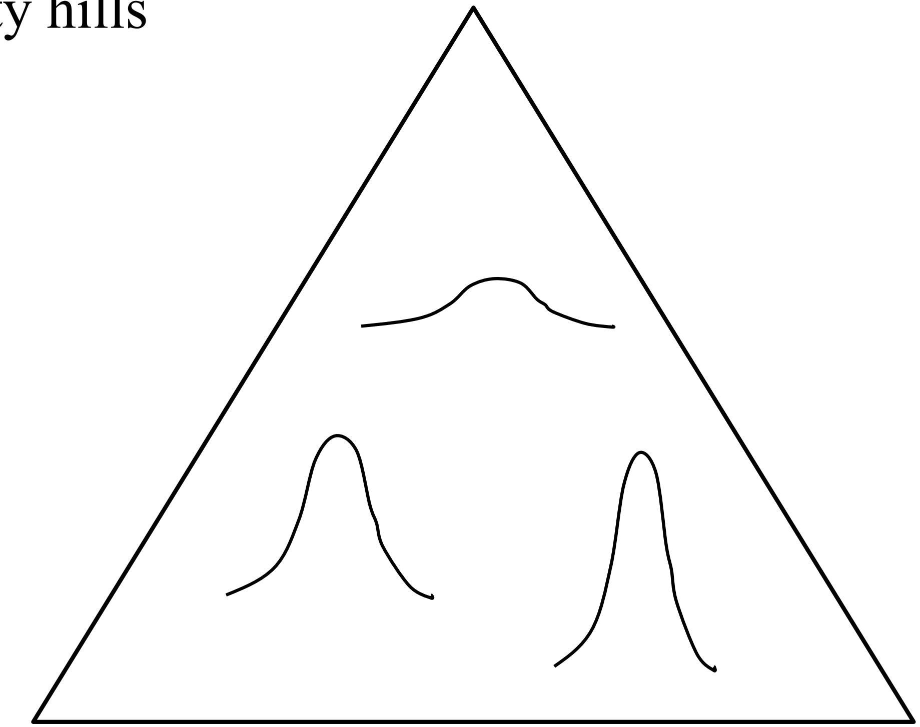


No one knows how to construct a DM prior from first principles.

Dirichlet Mixtures

The superposition of M *Dirichlet components*, with positive weights w_i that sum to 1, yielding a total of $M(L+1)-1$ free parameters.

We may visualize a Dirichlet mixture (DM) as a collection of probability hills in multinomial space.



No one knows how to construct a DM prior from first principles. So we invert the problem….

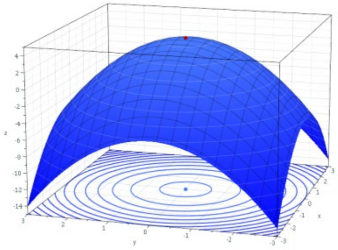
Given a set of properly aligned columns, what is the maximum-likelihood DM?

Optimization in High-Dimensional Space

Smooth and simple landscapes

Relatively easy and fast to find optimum.

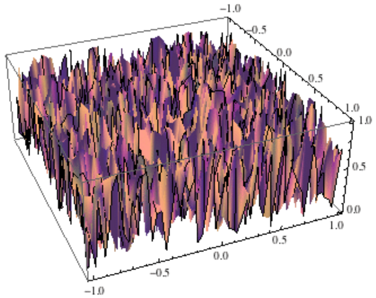
Algorithms: Newton's method; gradient descent.



Random landscapes

Finding optimal solution intractable.

Algorithms: Brute force enumeration.



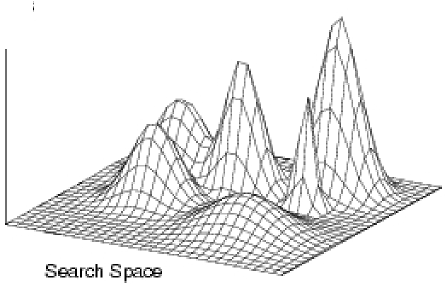
Rough but correlated landscapes

Difficult to find provably optimum solution.

Fairly effective heuristic methods available.

Algorithms: Simulated annealing; EM; *Gibbs sampling*.

Difficulties: Local optima.



Geman, S. & Geman, D. (1984) *IEEE Trans. Pattern Analysis and Machine Intelligence* **6**:721-741.

Gibbs Sampling for Dirichlet Mixtures

Given: A large set of multiple-alignment columns

Find: The M-component DM maximizing the likelihood of the data

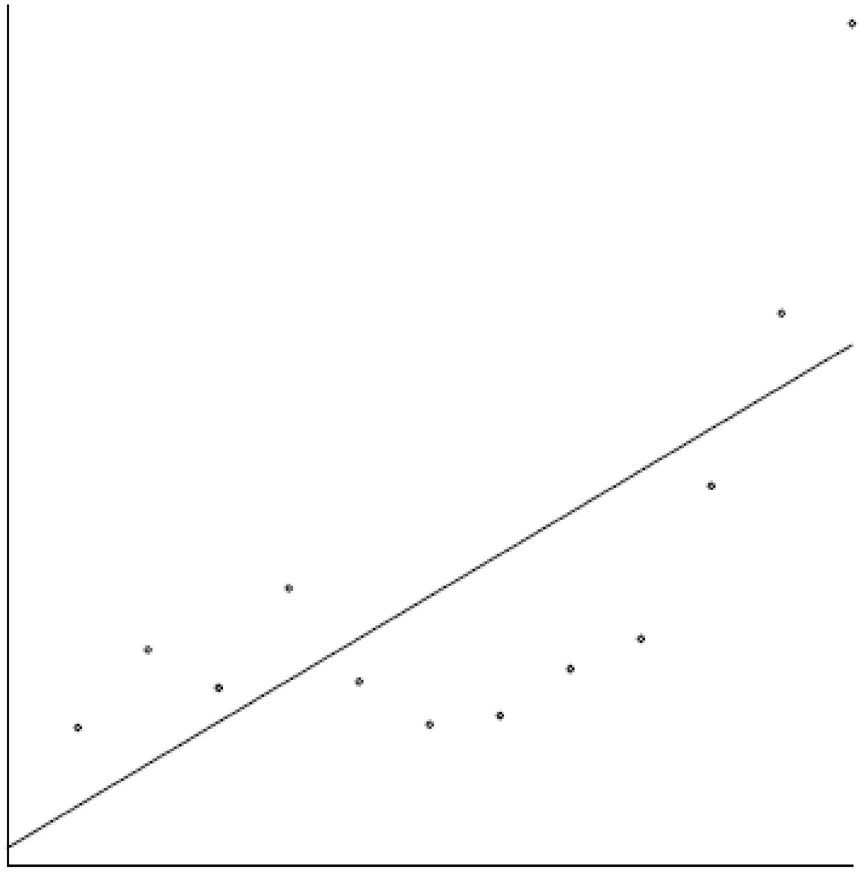
Algorithm

1) Initialize: Assign columns to components
2) Derive $(\vec{p}, \alpha, w)$ for each component from its columns
3) In turn, sample columns into new components, using probabilities proportional to implied likelihoods
4) Iterate

But: How many Dirichlet components should there be?

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

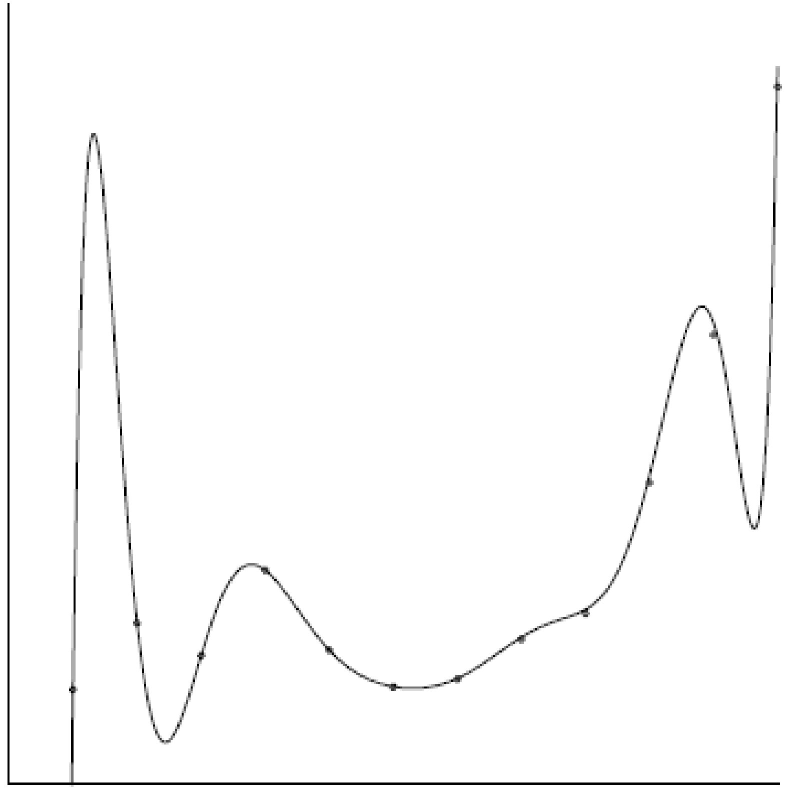
A model that is too simple underfits the data



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

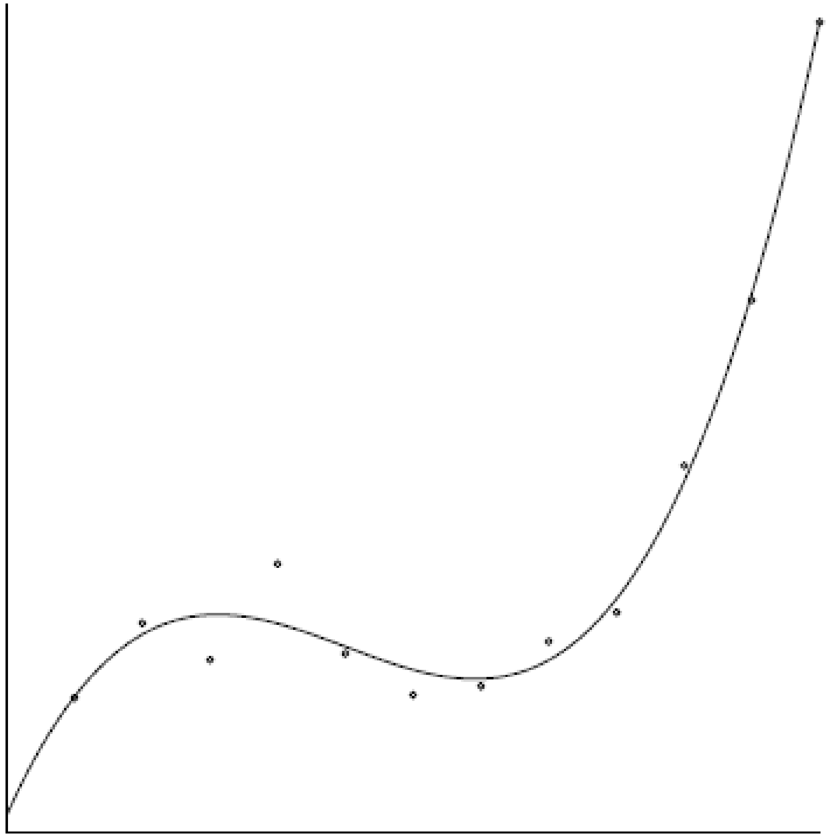
From: "A tutorial introduction to the minimum description length principle" by Peter Grünwald

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



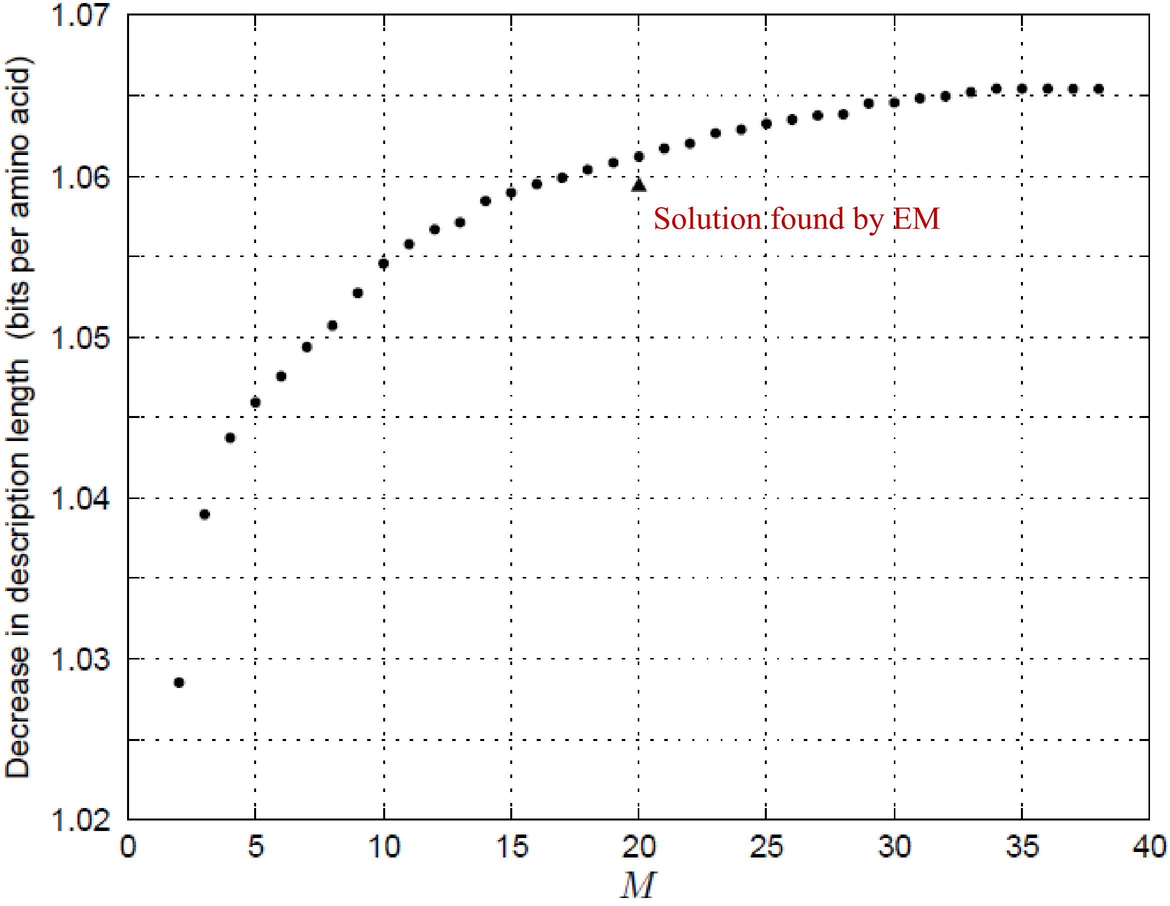
Everything should be made as simple as possible, but not simpler. – Albert Einstein

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

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

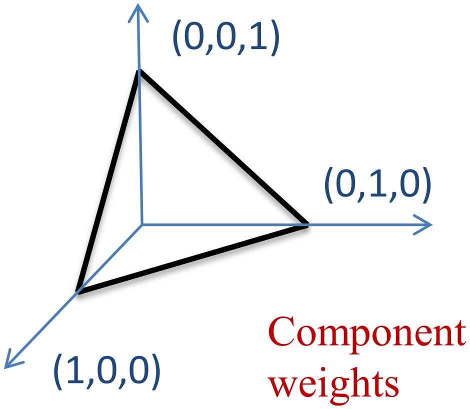
Problem: How effective is the algorithm at finding a maximum-likelihood DM?

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

The Dirichlet Process

The DP models *mixtures* of an *underlying distribution* with an unknown and unbounded number of components.

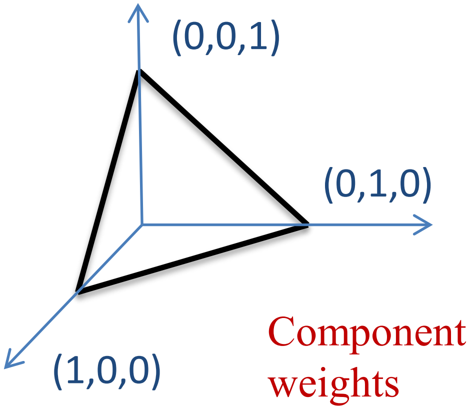
It generalizes the Dirichlet distribution to infinitely many dimensions, as a model of component *weights*.



The Dirichlet Process

The DP models *mixtures* of an *underlying distribution* with an unknown and unbounded number of components.

It generalizes the Dirichlet distribution to infinitely many dimensions, as a model of component *weights*.



A DP is specified by:

A prior *H* for the parameters of the underlying distribution

A parameter γ defining a prior for the component weights

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

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

The Chinese Restaurant Process

A restaurant with infinitely many tables, which can each seat infinitely many people.

As people enter, they sit at tables randomly, but prefer company:

They choose occupied tables with probability proportional to the number of people already seated there;

They choose a new, unoccupied table, with probability proportional to γ.

The Chinese Restaurant Process

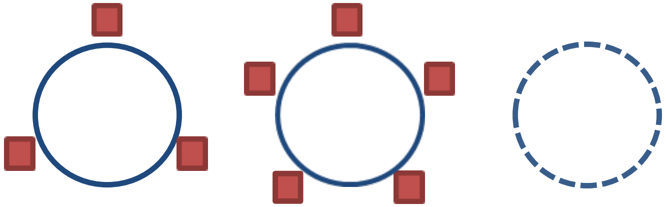
A restaurant with infinitely many tables, which can each seat infinitely many people.

As people enter, they sit at tables randomly, but prefer company:

They choose occupied tables with probability proportional to the number of people already seated there;

They choose a new, unoccupied table, with probability proportional to γ.

Example: 8 people already seated; $\gamma = 2$



Probability: 0.3 0.5 0.2

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

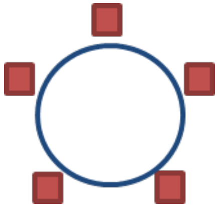
Dirichlet-Process Modifications to Gibbs Sampling

When sampling a column C into a component:

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

Allow C to seed a new component, with probability proportional to γ:

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



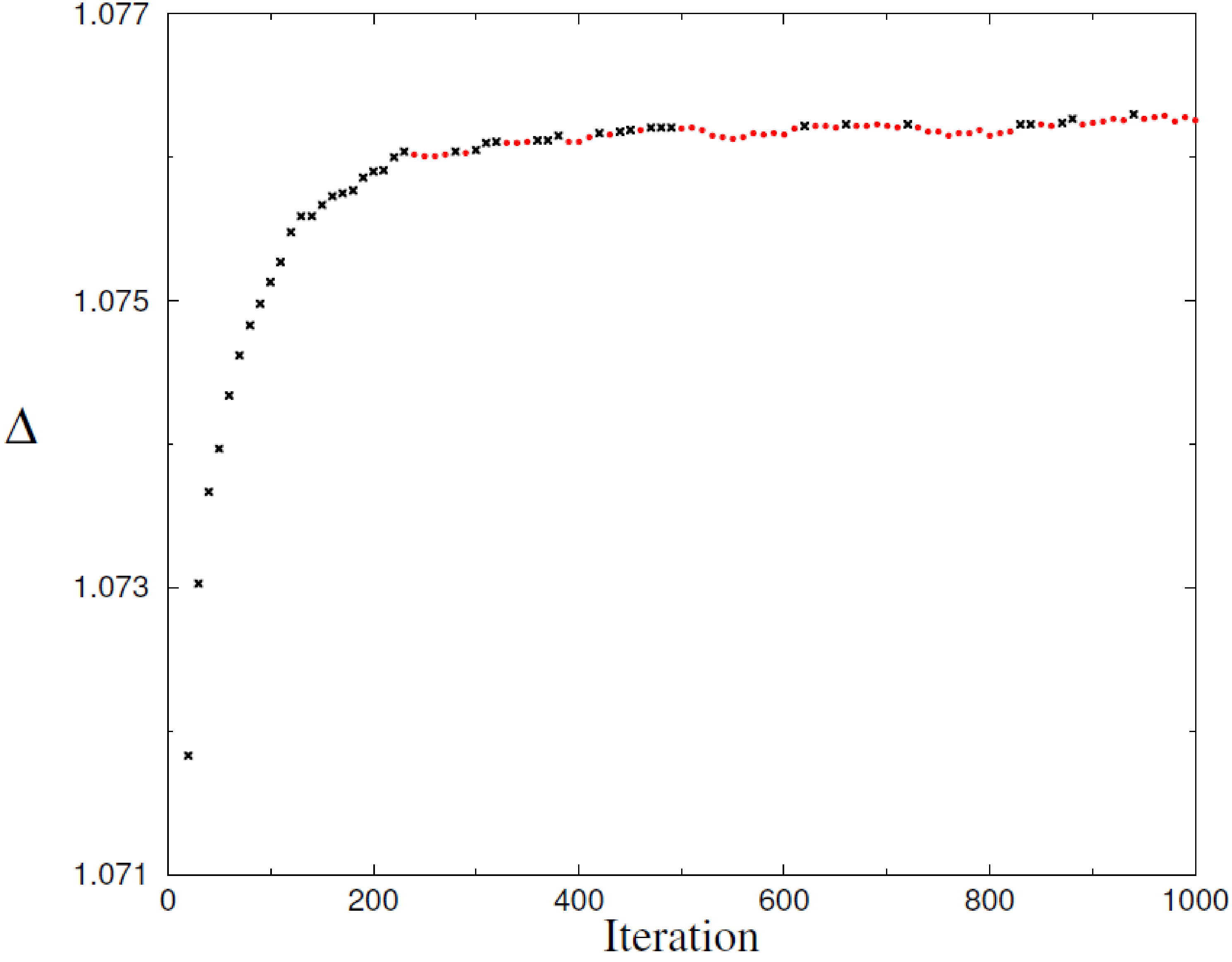
$$\text{Prob(new component)} \propto \gamma \frac{\Gamma(\beta)}{\Gamma(\beta + c)} \prod_{j=1}^{20} \frac{\Gamma(\beta p_j + c_j)}{\Gamma(\beta p_j)}$$

To calculate Dirichlet parameters for the component:

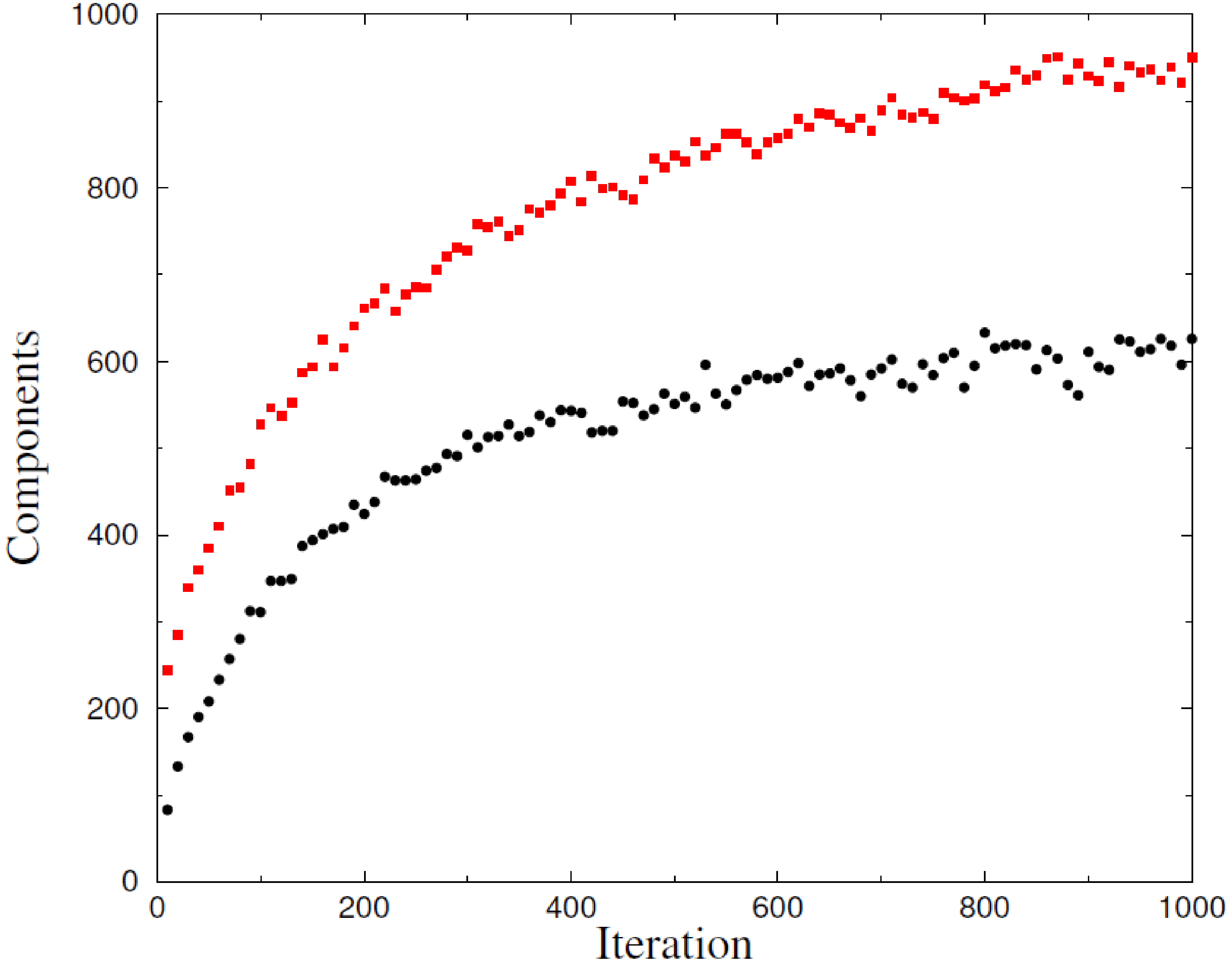
Sample from the posterior implied by H and the component’s columns.

Nguyen, V.-A., *et al.* (2013) *J. Comput. Biol.* **20**:1-18.

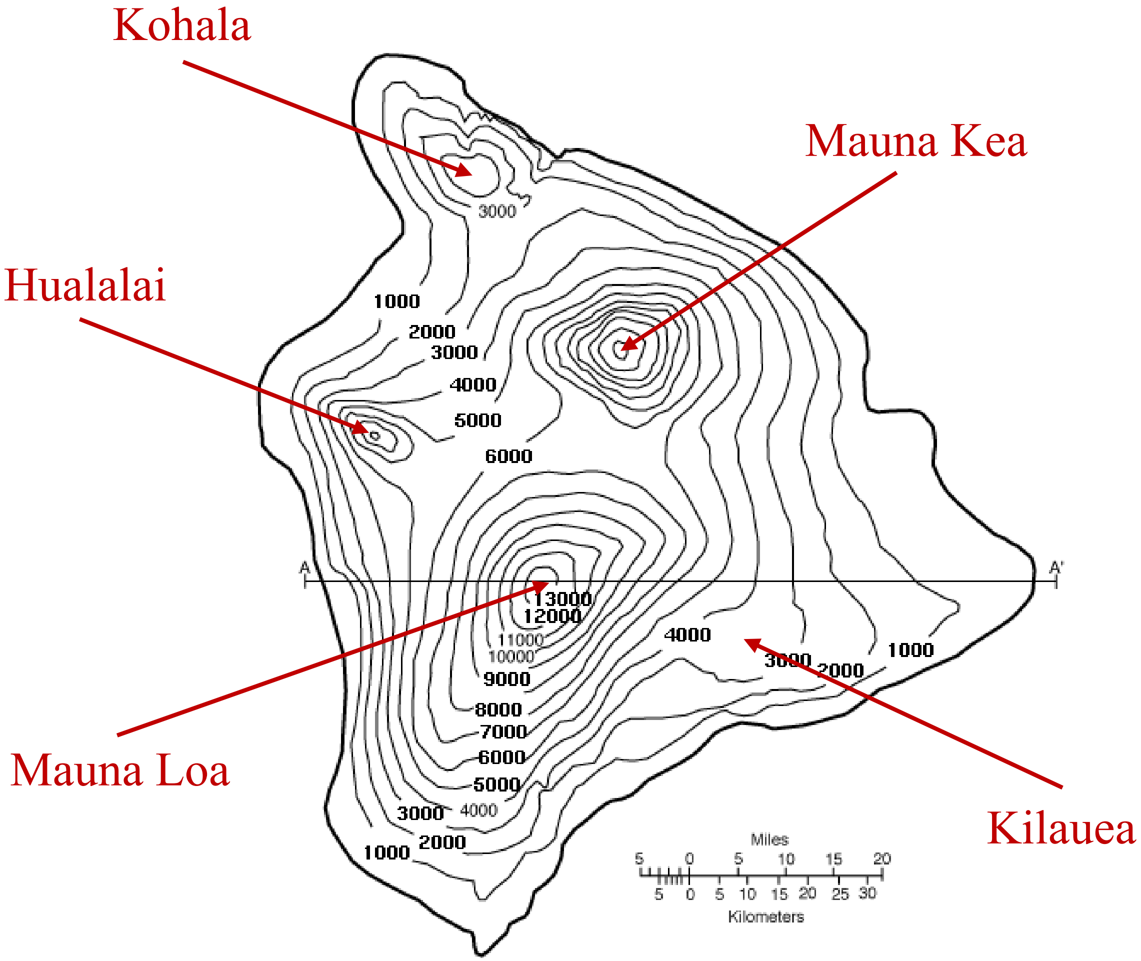
Decrease in Total Description Length as a Function of DP-Sampler Iteration ($\beta = 400$; $\gamma = 100$)



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



Topographic Map of the Big Island of Hawai'i



Kohala

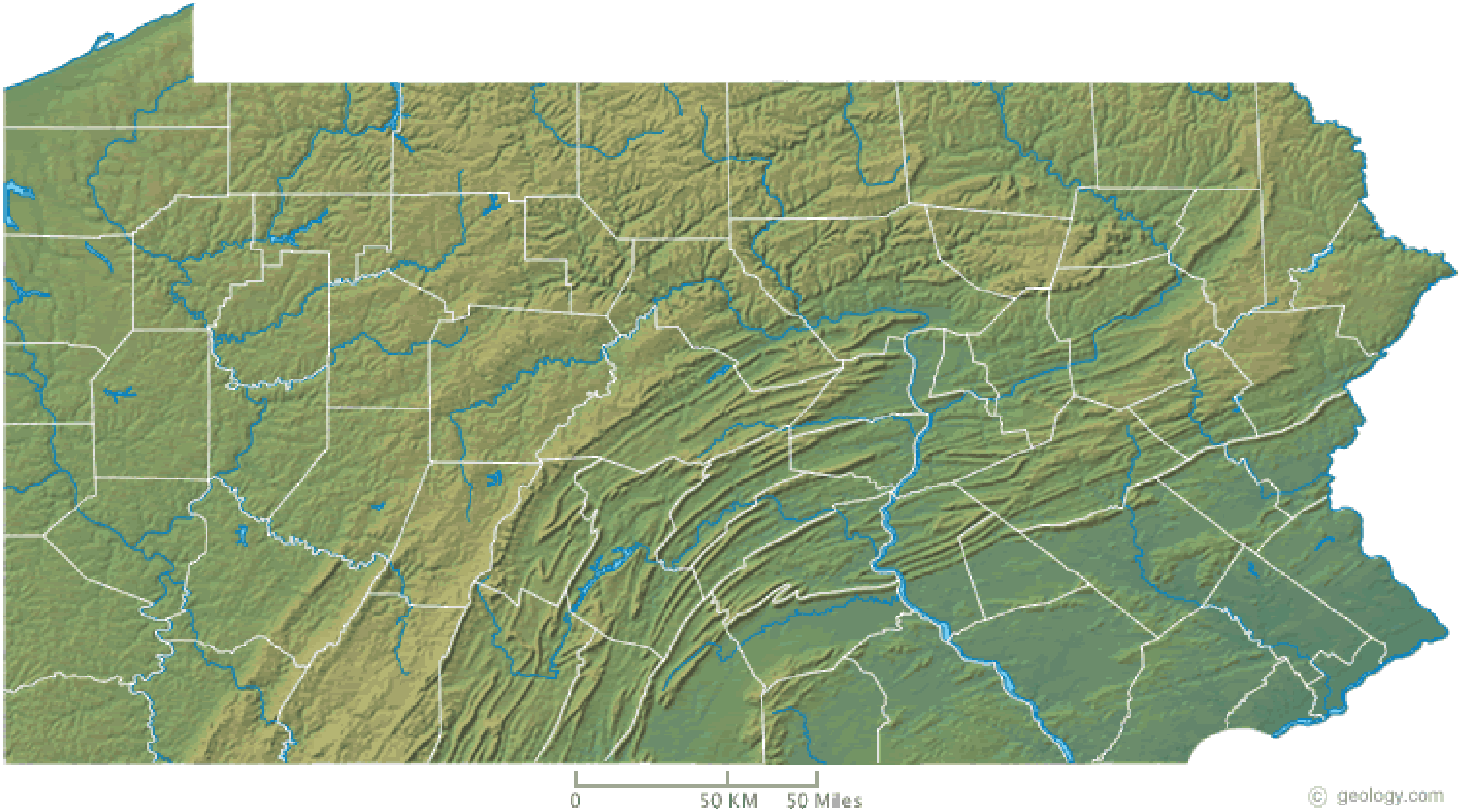
Mauna Kea

Hualalai

Mauna Loa

Kilauea

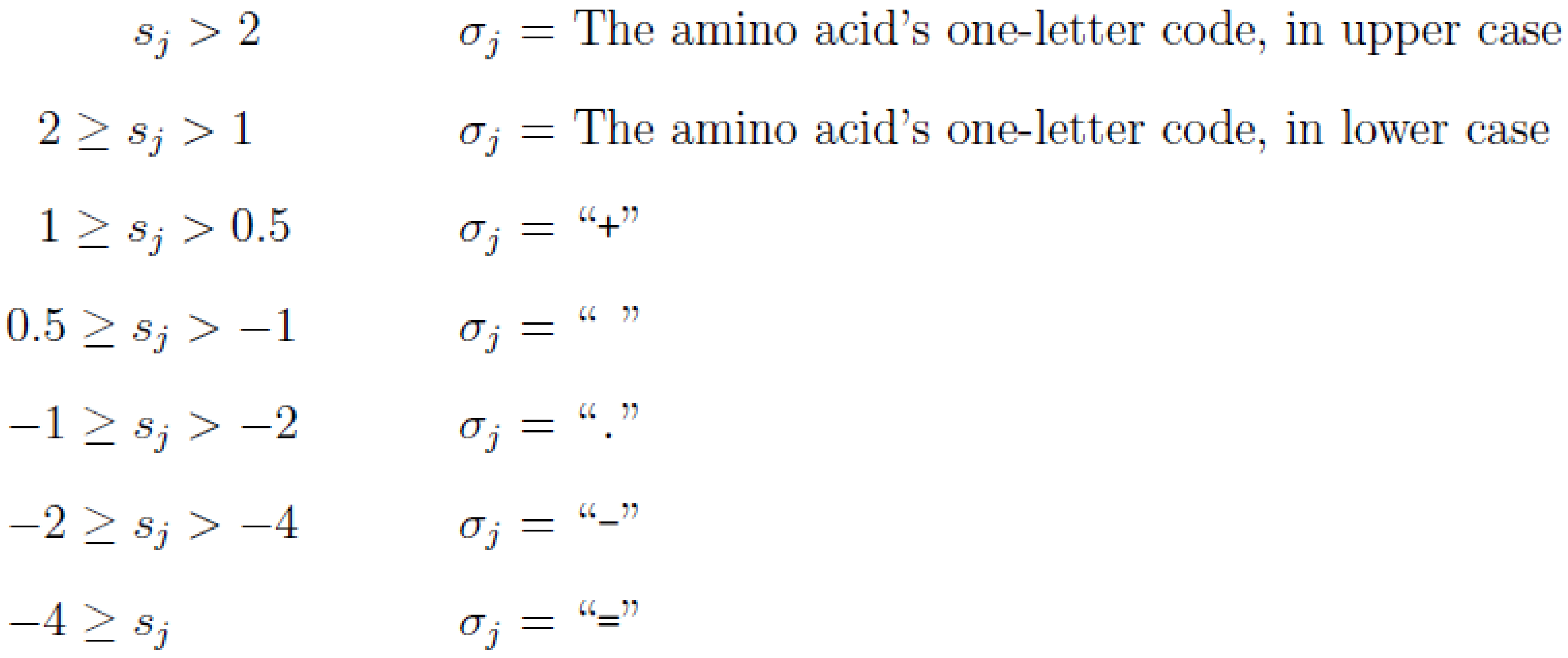
Topographic Map of Pennsylvania



Visualizing Dirichlet Mixture Components

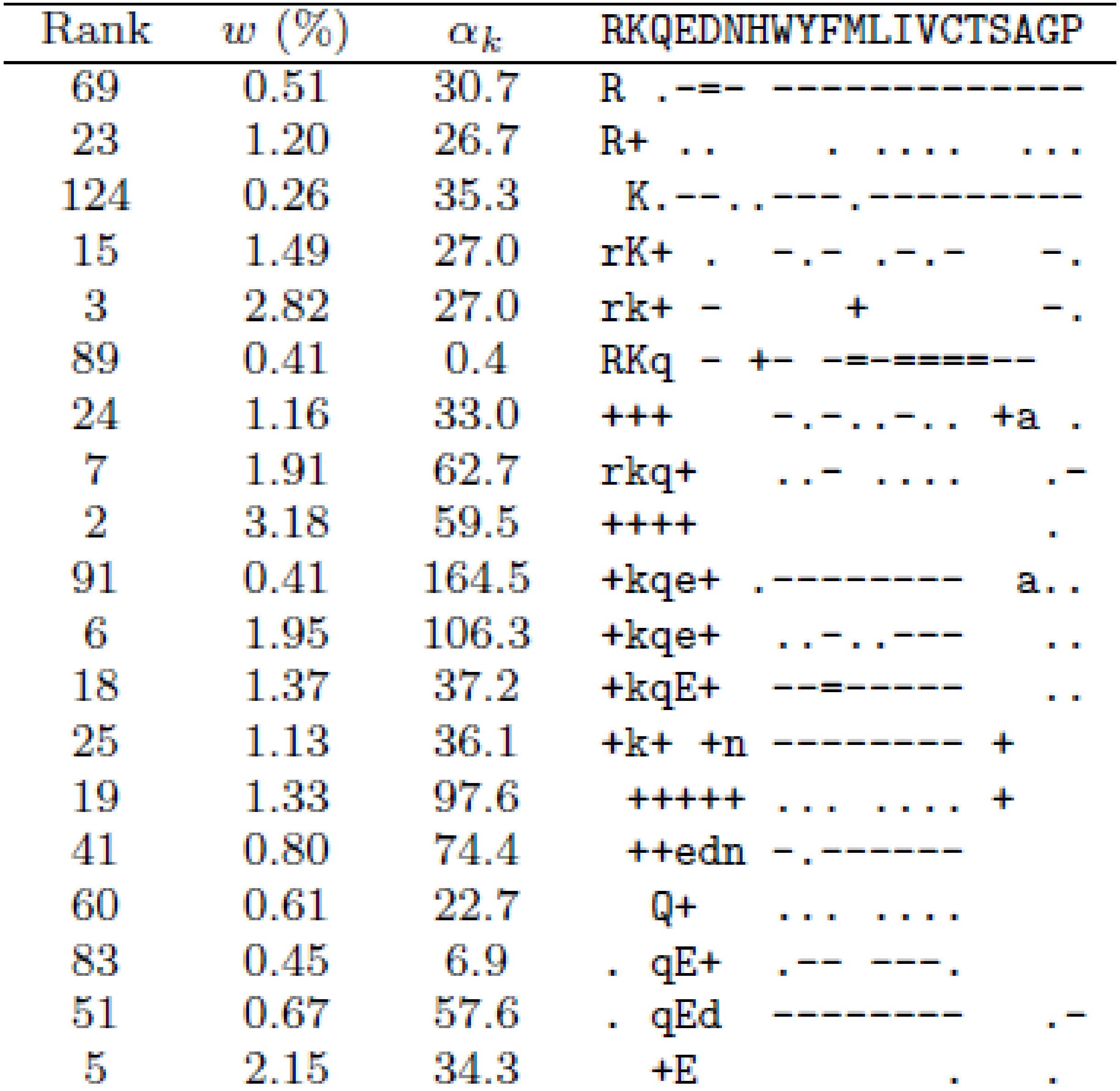
Reorder the amino acids: RKQEDNHWYFMLIVCTSAGP

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



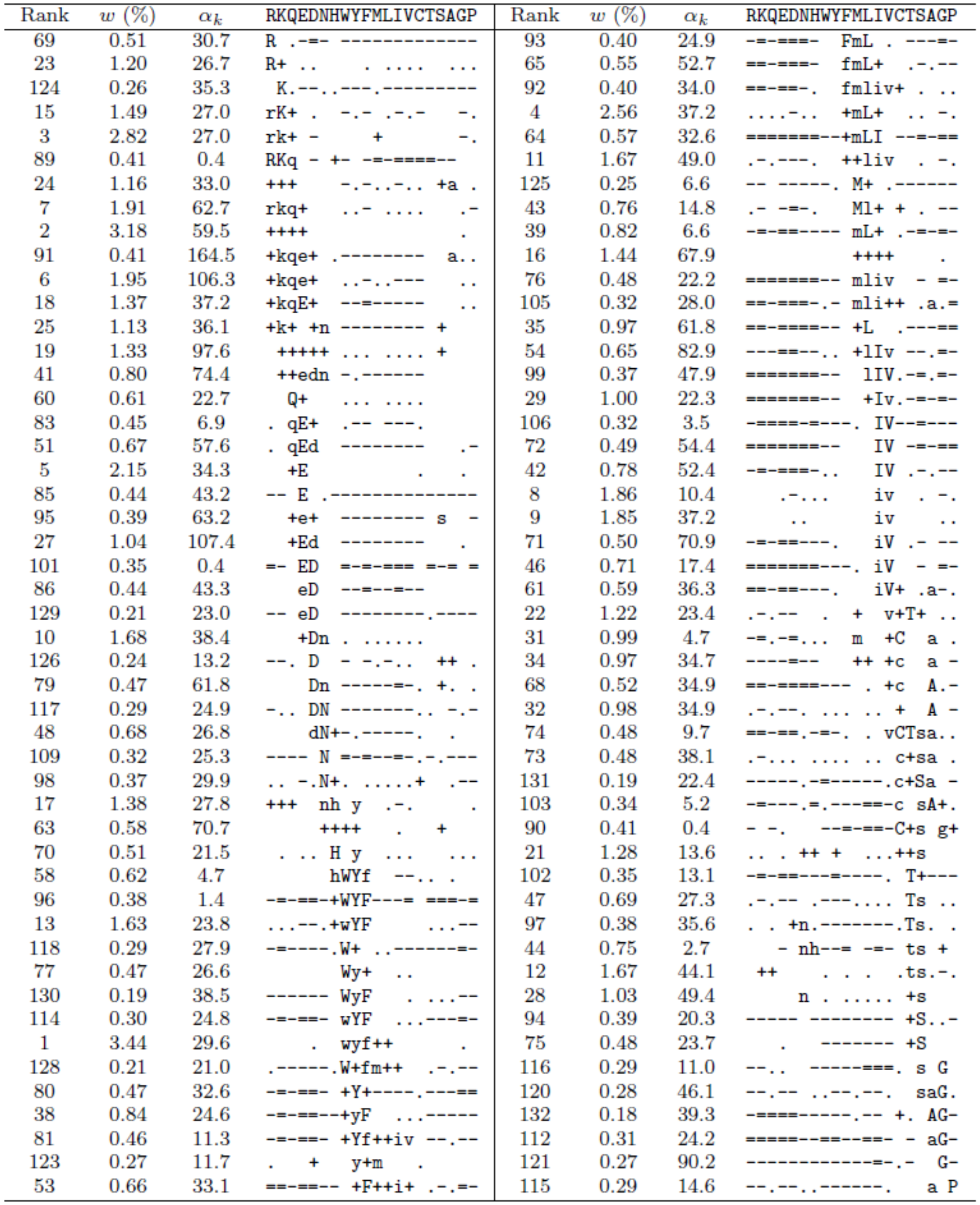
$s_j > 2$	$\sigma_j =$ The amino acid's one-letter code, in upper case
$2 \geq s_j > 1$	$\sigma_j =$ The amino acid's one-letter code, in lower case
$1 \geq s_j > 0.5$	$\sigma_j =$ "+"
$0.5 \geq s_j > -1$	$\sigma_j =$ " "
$-1 \geq s_j > -2$	$\sigma_j =$ "."
$-2 \geq s_j > -4$	$\sigma_j =$ "–"
$-4 \geq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

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     ...--
```

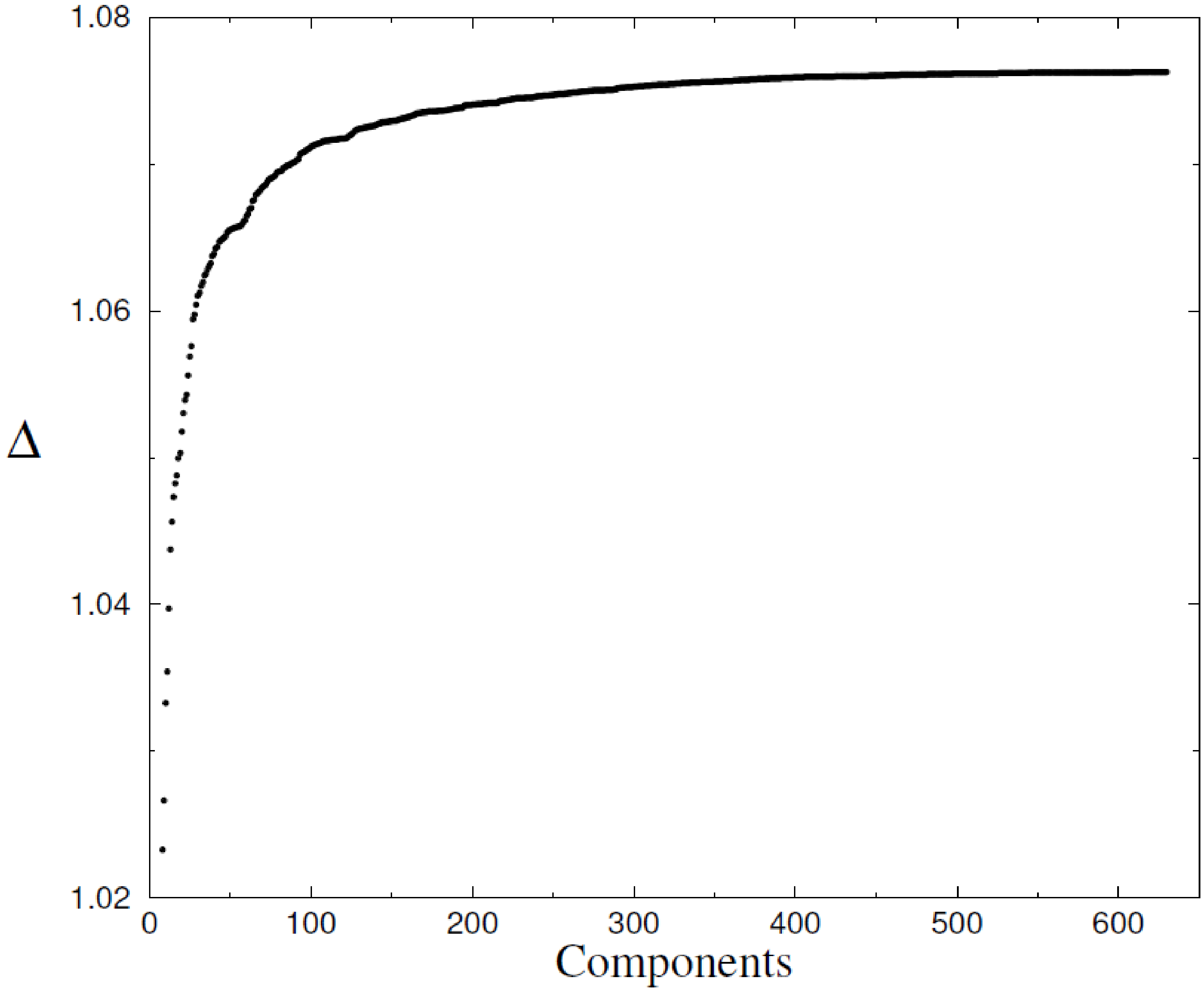
Group B: Hydrophylic Positions Favoring Glycine or Proline

Rank	w (%)	α	R	K	Q	E	D	N	H	W	Y	F	M	L	I	V	C	T	S	A	G	P
100	0.36	32.5	+	k	+			n		-	-	-	-	-	=	-	-	.		.	G	.
82	0.46	38.1	.				d	n		-	.	-	-	-	=	-	-	.		.	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								-	-	-	-	-	-	-	-				g	P
45	0.72	75.9				+	d	+		.	.	-	-	-	-	-	.		+		+	p
108	0.32	31.7	.				d	+		-	-	-	-	-	-	-	.		s			P
127	0.21	77.4					d	+		-	-	-	-	-	-	-	.	t	s	.		p
56	0.63	69.9				e	d			-	-	-	-	-	-	-	-					P
110	0.31	84.8	+	k	+	e	+			-	-	-	-	-	-	-	-					p
119	0.28	9.2	r	k			d	+		-	-	=		-	=	-	-			.		p
50	0.67	41.6	r	k	+					-	.	-	.	.	.	.	.					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

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



Collaborators

National Center for Biotechnology Information

Xugang Ye
Yi-Kuo Yu

University of Maryland, College Park

Viet-An Nguyen
Jordan Boyd-Graber