Phylogenetic modeling and the Dirichlet process

Adam J. Richards

Centre national de la recherche scientifique (CNRS)
(French National Center for Scientific Research)
Station d'Ecologie Expérimentale du CNRS à Moulis

Last updated: October 16, 2014

Codon substitution models

There are $4 \times 4 \times 4 = 64$ possible codons. 61 code for amino acids while the 3 others are stop codons. So most amino acids are encoded by more than on codon allowing for substitutions in the genetic code that do not change the amino acid sequence (synonymous) substitutions.

- A major focus has been put on applying a mechanistic approach rather than phenomenological one [Rodrigue and Philippe, 2010]
- Generative models!
- Site-heterogeneity i.e. [Rodrigue et al., 2010]

Dirichlet Process

A stochastic process used in Bayesian nonparametric models of data. It is a distribution over distributions, where each draw from a Dirichlet process is itself a distribution

- Parametric function estimation (e.g. regression, classification)
- Nonparametric function estimation with Gaussian Processes
- Parametric density estimation (e.g. Gaussian mixture models)
- Bayesian nonparametric density estimation with DP
- Semiparametric modeling (e.g. GLMMs but with nonparametric noise and/or nonparametric random effects)
- Model selection/averaging (clustering, neuron spike sorting, topic modeling, computer vision)

See a lecture on Dirichlet processes by Yee Whye Teh for individual references

Dirichlet Process (DP)

Also known as the Ferguson distribution [Ferguson, 1973]. DP is a distribution over distributions and was motivated by Bayesian density estimation.

Suppose H is a probability distribution, and G is a random probability distribution, both with support in space $\mathbb{X}$. Then G is distributed according to a DP with base distribution H, and precision parameter $\alpha > 0$, for all finite and measurable partitions of $\mathbb{X}$.

A Dirichlet distribution is a distribution over the K-dimensional probability simplex

$$\Delta_K = \{(\pi_1, \ldots, \pi_K) : \pi_k \geq 0, \sum_k \pi_k = 1\} \quad (1)$$

$(\pi_1, \ldots, \pi_K)$ is Dirichlet distributed, i.e.

$$(\pi_1, \ldots, \pi_K) \sim \text{Dirichlet}(\alpha_1, ..., \alpha_K)$$

with parameters $(\alpha_1, ..., \alpha_K)$, if

$$p(\pi_1, \ldots, \pi_K) = \frac{\Gamma(\sum_k \alpha_k)}{\sum_k \Gamma(\alpha_k)} \prod_{k=1}^{n} \pi_k^{\alpha_k - 1}$$

Properties of Dirichlet distributions

- Agglomerative property

$$(\pi_1, \ldots, \pi_K) \sim \text{Dirichlet}(\alpha_1, ..., \alpha_K)$$
$$(\pi_1 + \pi_2, \pi_3, \ldots, \pi_K) \sim \text{Dirichlet}(\alpha_1 + \alpha_2, \alpha_3, ..., \alpha_K)$$

  Also, works for partitions of π_i

- Decimative property

$$(\pi_1, \ldots, \pi_K) \sim \text{Dirichlet}(\alpha_1, ..., \alpha_K)$$
$$(\pi_1\tau_1, \pi_2\tau_2, \ldots, \pi_K) \sim \text{Dirichlet}(\alpha_1\beta_1, \alpha_2\beta_2, \ldots, \alpha_K)$$

DP parameters

A DP has two parameters

- Base distribution H - like the mean of the DP
- Strength parameter α - like an inverse-variance of the DP

$$G \sim \text{DP}(\alpha, H)$$

And for any partition $(A_1, \ldots, A_K)$ of $\mathbb{X}$:

$$(G(A_1), \ldots, G(A_k)) \sim \text{Dirichlet}(\alpha H(A_1), \ldots, \alpha H(A_K))$$

Note that the H is sometimes referred to as G_0.

Polya Urn Scheme

Let $\boldsymbol{\theta} = \{\theta_1, \ldots, \theta_N\}$ be a sequence of random variables. Drawn independently from a DP

$$\theta_j \sim G$$
$$G \sim \text{DP}(\alpha, H)$$

Then the posterior of G is a DP with precision $\alpha + N$ and base distribution

$$\frac{\alpha}{\alpha + N} H + \frac{1}{\alpha + N} \sum_{j=1}^{N} \delta_{\theta_j}$$

where δ_θ Dirac probability mass function, placing all mass at θ. Given this α is interpreted as a *prior sample size*, or the strength of prior belief in the base measure H. Polya urn scheme - generative construction for $\boldsymbol{\theta}$ identified by marginalizing w.r.t. G. The scheme yields:

$$p(\theta_j | \boldsymbol{\theta}_{1:j-1}) \propto \alpha H(\theta_j) + \sum_{k=1}^{j-1} \delta_{\theta_k} \tag{2}$$

Eqn 2 is positive where θ_j is the same as θ_k and they are clustered when the probabilities are identical.

[Blackwell and MacQueen, 1973]

Chinese restaurant process

- Draw $\theta_1, \ldots, \theta_N$ from a Blackwell-MacQueen urn scheme
- They take on $K < N$ distinct values, say $\theta_1^*, \ldots, \theta_K^*$
- This defines a partition of $1, \ldots, N$ into K clusters, such that if i is in cluster k , then $\theta_i = \theta_k^*$.
- Random draws $\theta_1, \ldots, \theta_N$ from a Blackwell-MacQueen urn scheme induce a random partition of $1, \ldots, N$.
- The induced distribution over partitions is a Chinese restaurant process.

[Aldous, 1985]

More, more, more

There are more details... but instead here are the appropriate references.

- DP - introduced by [Ferguson, 1973], while [Antoniak, 1974] further developed DPs and introduced mixtures of DPs.
- Blackwell-MacQueen urn scheme - [Blackwell and MacQueen, 1973] showed the scheme is exchangeable.
- Chinese restaurant process see [Aldous, 1985]
- MCMC is the primary means to summarize posterior quantities in DP models see [MacEachern, 1994, Escobar and West, 1995]
- An alternative representation of the DP is the stick-breaking construction [Sethuraman, 1994]
- Hierarchical Dirichlet Processes were first developed by [Teh et al., 2006]

But why the DP?

Cardinality

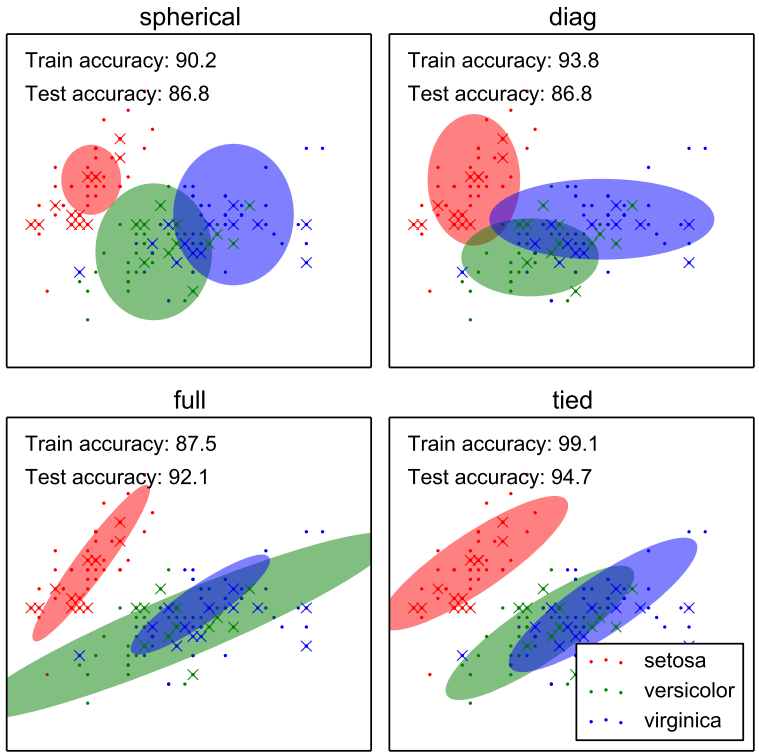
Particularly an issue for clustering problems. Although model selection and cross-validation are viable approaches to determine cardinality there are situations where a solution with these methods is generally intractable

- Bayesian model averaging - does not specify cardinality but rather average over several
- i.e. use a prior over the set of possible K clusters and let the data define a posterior
- difficult to interpret and computational limitations

Gaussian mixture model (GMM)

A probabilistic model that assumes data is generated from a mixture of a finite number of Gaussian distributions with unknown parameters.

- assumes a covariance structure i.e. spherical, diagonal, full, and tied
- full covariance normally performs best though it is overfits on small datasets
- plot: iris dataset — training data (dots), test data (crosses)



Modified from scikit-learn documentation

GMMs continued

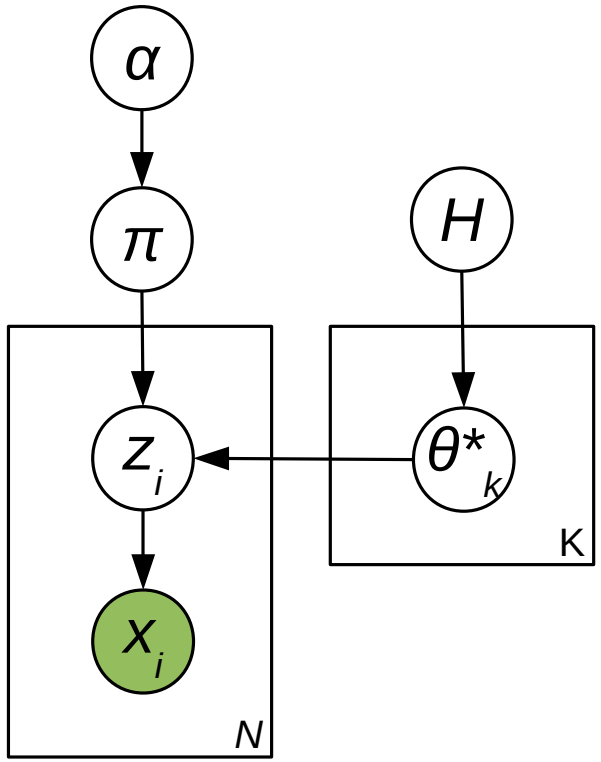
Different forms of inference exist:

- Expectation-Maximization (EM)
  - fast
  - singularities (infinite likelihood)
  - need to specify the number of components
- Variational inference
  - avoids singularities so we can use full covariance in high dimensions
  - will bias all means towards the origin and covariances tend towards spherical
  - need to specify hyperparameter (cross-validation)
- MCMC
  - exact solution upon convergence
  - computational costly

Finite mixture models

$$\theta_k^* \sim H$$
$$\pi \sim \text{Dirichlet}(\alpha/K, \ldots, \alpha/K)$$
$$z_i|\pi \sim \text{Discrete}(\pi)$$
$$x_i|\theta_{z_i}^* \sim F(\cdot|\theta_{z_i}^*)$$

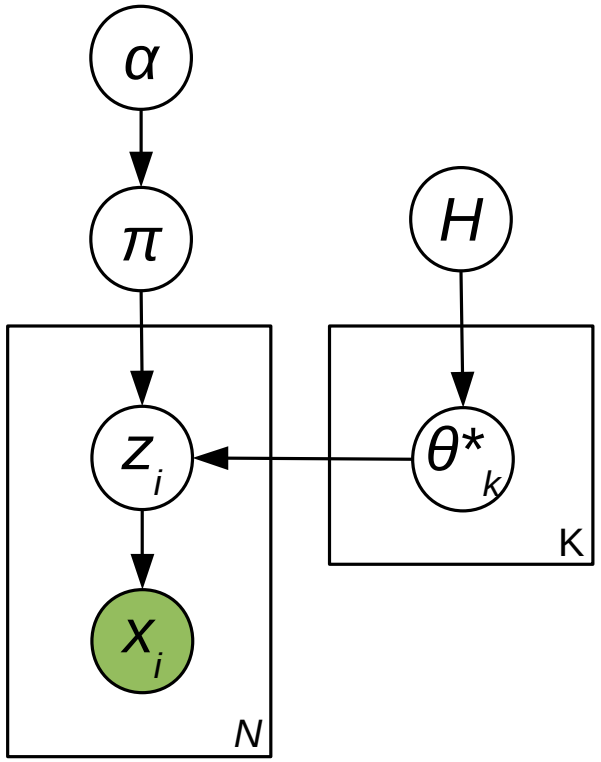
Still have to use model selection/averaging over the hyperparameters in H, the Dirichlet parameter α and the number of components K



Infinite mixture models

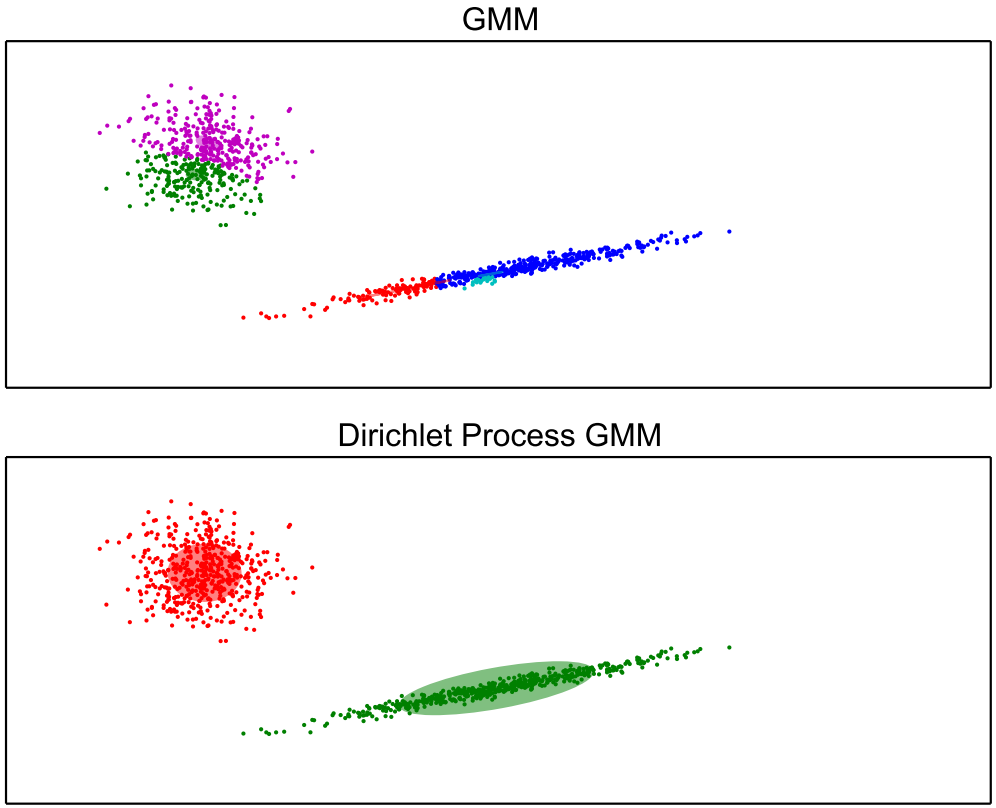
Dirichlet Process Gaussian Mixture Model (DPGMM) — an infinite mixture model with the Dirichlet Process as a prior distribution on the number of clusters.

- Let K be very large
- If parameters θ_{k*} and mixing proportions π integrated out, the number of latent variables left does not grow with K and we have no overfitting
- At most N components will be associated with the data



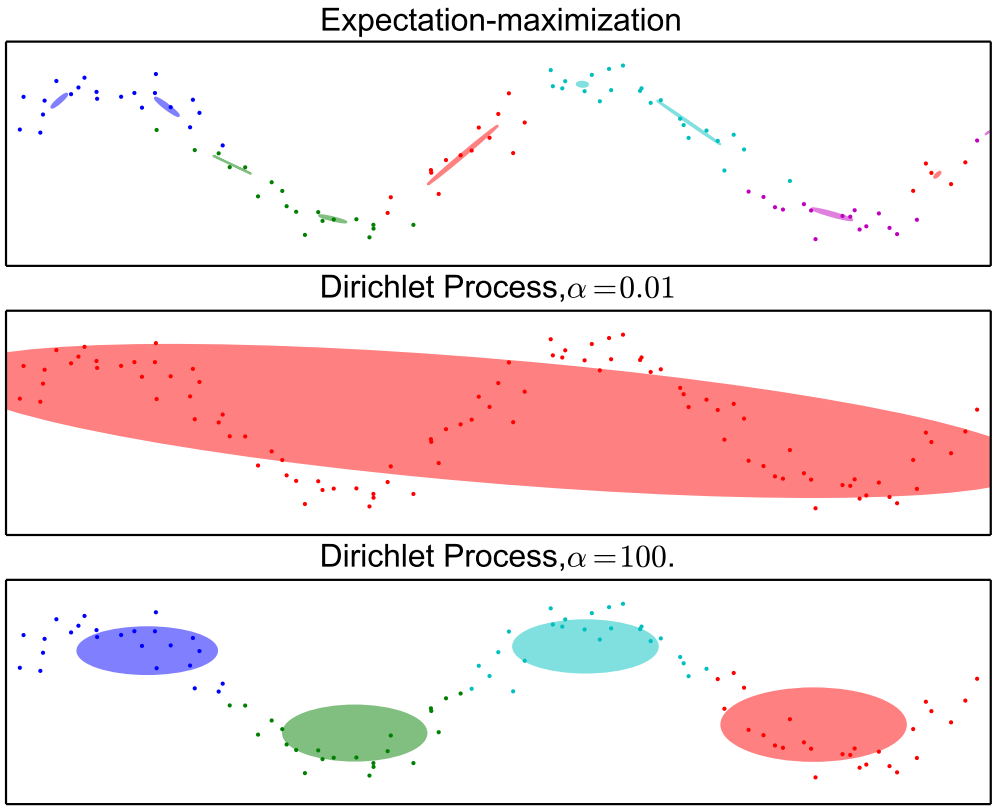
The Infinite Gaussian mixture model [Rasmussen, 2000]

If we specify $K = 5$...



Modified from scikit-learn documentation. This model is implemented using variational inference as derived in [Blei and Jordan, 2005]

a closer look at α



The strength parameter α acts like the precision (inverse variance) of the DP

Modified from scikit-learn documentation

CAT model

- Relax the assumption that proteins evolve under the sample substitution process (20×20 substitution matrix)
- AA replacement at different sites of a protein alignment can have distinct substitution processes
- CAT model assumes distinct processes (classes) differing by equilibrium frequencies over the 20 residues
- Using a DP the affiliations of each site to a given class are free variables
- Substitutional heterogeneity is estimated using posterior means (classes)
- Data come in the form on an alignment of P amino acid sequences of length N.
- Substitutions occur according to a rate matrix Q_{lm} expressed in terms of 20 probabilities or equilibrium frequencies.

[Lartillot and Philippe, 2004]

CAT continued

Let π_l be the set of 20 equilibrium frequencies, s.t. $\sum_{l=1}^{20} = 1$. And let ρ_{lm} be the exchangeability parameters that are assumed to hold the relation,

$$Q_{lm} = \frac{1}{Z}\rho_{lm}\pi_m, l \neq m$$
$$Q_{ll} = -\sum_{m \neq l} Q_{lm}$$

The process is assumed to be reversible $Q_{lm} = Q_{ml}$ and the matrix is scaled to 1 using the normalizing constant

$$Z = 2 \times \sum_{1 \leq l \leq m \leq 20} \rho_{lm}\pi_l\pi_m \tag{3}$$

[Lartillot and Philippe, 2004]

CAT continued

Branch lengths are measured as the expected number of substitutions per site. From Q the transition probability matrix $P(v) = [P_{lm}(v)]$ can be used to specify the probability that amino-acid l changes into m over an evolutionary distance of v via $P(v) = e^{vQ}$

Under the CAT model sites are distributed according to a mixture of K distinct classes– each class is characterized by its own substitution matrix Q^k. An classes are specified using the vector z.

[Lartillot and Philippe, 2004]

Muse and Gaut Model (MG)

The following formulation, inspiried by Muse and Gaut, is a basic codon substitution model that has two sets of parameters.

$$\rho = (\rho_{lm}) \text{ where } lm \in \{1 \dots 4\} \text{ and } \sum \rho_{lm} = 1 \tag{4}$$

$$\varphi = (\varphi_m) \text{ where } m \in \{1 \dots 4\} \text{ and } \sum \varphi_m = 1 \tag{5}$$

$$Q_{ab} = \begin{cases} \rho_{a_c b_c}, & \text{if a and b are synonymous and differ only at } c^{\text{th}} \text{codon position} \\ \omega \rho_{a_c b_c} \varphi_{b_c}, & \text{if a and b are nonsynonymous and differ only at } c^{\text{th}} \text{codon position} \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

a_c corresponds to the index of the nucleotide at the c^{th} ($c \in \{1, 2, 3\}$) position of codon a. Extensions to the model model also includes ω, modulating nonsynonymous rates without regard to the amino acids involved (MG-NS) or instead of ω we can apply a Dirichlet process approach to capture across-site heterogeneity in nonsynonymous mutation rates (MG-NSDP).

[Muse and Gaut, 1994]

Yang and Nielsen (CP)

This approach, inspired by Yan and Nielsen, uses a set of 61 codon fitness parameters.

$$\psi = (\psi_a) \text{ where } a \in \{1 \ldots 61\} \text{ and } \sum a = 1 \tag{7}$$

$$Q_{ab} = \begin{cases} \rho_{a_c b_c} \varphi_{b_c} \left(\frac{\psi_b}{\psi_a}\right)^{1/2}, & \text{if a and b are synonymous and differ only at } c^{\text{th}} \text{ codon position} \\ \omega \rho_{a_c b_c} \varphi_{b_c} \left(\frac{\psi_b}{\psi_a}\right)^{1/2}, & \text{if a and b are nonsynonymous and differ only at } c^{\text{th}} \text{ codon position} \\ 0, & \text{otherwise.} \end{cases} \tag{8}$$

With or without ω the models are referred to as Codon Preference (CP). Therefore we have MG-CP, MG-NS-CP, and MG-NSDP-CP.

[Yang and Nielsen, 2008]

Robinson *et al.* (SC)

Approach inspired by Robinson *et al.* Rodrigue *et al.* define a model in sequence space. Rates are given from one sequence state s to another s.

$$R_{SS'} = \begin{cases} \rho_{s_{i_c} s'_{i_c}} \varphi_{s_{i_c}}, & \text{if A} \\ \omega \rho_{s_{i_c} s'_{i_c}} \varphi_{s_{i_c}} e^{\beta(G_{(s)} - G_{(s')})}, & \\ & \text{if B} \\ 0, & \text{otherwise.} \end{cases} \tag{9}$$

where s_i is the codon at the i^{th} site of sequence s and s_{i_c} is the nucleotide at the c^{th} codon of the i^{th} site of sequence s.

- A s and s' differ only only at c^{th} codon position at the i^{th} site (implies synonymous change)
- B s and s' differ only only at c^{th} codon position at the i^{th} site (implies nonsynonymous change)

For a given sequence s, $G(s)$ returns a pseudo-energy score of sequence-structure compatibility (see [Kleinman et al., 2006]). These are referred to as Structurally Constrained (SC) models. MG-SC, MG-NS-SC and MG-NSDP-SC.

[Robinson et al., 2003, Rodrigue et al., 2009]

SC + CP

$$\psi = (\psi_{i_c}) \text{ where } i_c \in \{1 \ldots 61\} \text{ and } \sum i_c = 1 \tag{10}$$

$$R_{SS'} = \begin{cases} \rho_{s_{i_c} s'_{i_c}} \varphi_{s_{i_c}} \left(\frac{\psi_{s'_i}}{\psi_{s_i}} \right)^{1/2}, & \text{if A} \\ \omega \rho_{s_{i_c} s'_{i_c}} \varphi_{s_{i_c}} \left(\frac{\psi_{s'_i}}{\psi_{s_i}} \right)^{1/2} e^{\beta(G_{(s)} - G_{(s')})}, & \text{if B} \\ 0, & \text{otherwise.} \end{cases} \tag{11}$$

where s_i is the codon at the i^{th} site of sequence s and s_{i_c} is the nucleotide at the c^{th} codon of the i^{th} site of sequence s.

- A s and s' differ only only at c^{th} codon position at the i^{th} site (implies synonymous change)
- B s and s' differ only only at c^{th} codon position at the i^{th} site (implies nonsynonymous change)

These models are referred to as MG-CP-SC, MG-NS-CP-SC, and MG-NSDP-CP-SC.

[Rodrigue and Philippe, 2010]

Model comparisons

Table I. Natural logarithm of the Bayes factor for models considered, with MG-NS used as a reference[a]

Model	β-globin	*adh*
MG	[-92.0; -91.8]	[-319.1; -316.3]
MG-SC	[-22.3; -21.8]	[-220.8; -217.7]
MG-CP	[-4.4; -1.8]	[-218.3; -211.9]
MG-CP-SC	[83.1; 89.2]	[-130.8; -120.9]
MG-NS	-	-
MG-NS-SC	[48.5; 49.5]	[58.1; 58.6]
MG-NS-CP	[122.1; 123.8]	[165.5; 168.7]
MG-NS-CP-SC	[184.8; 188.3]	[225.6; 235.2]
MG-NSDP	[102.2; 104.2]	[96.8; 100.3]
MG-NSDP-SC	[185.7; 188.4]	[177.7; 181.6]
MG-NSDP-CP	[236.5; 241.0]	[254.1; 265.3]
MG-NSDP-CP-SC	[316.4; 321.5]	[328.0; 342.8]

[a]Values given are the upper and lower consistency checks from bi-directional thermodynamic integrations (giving a crude sense of computational error). We used the thermodynamic integration methods described in Refs [6,10] to perform a model contrast via Bayes factor calculation for two datasets studied in these last two works: 17 vertebrate sequences of the β-globin gene, and 23 *D. melanogaster* sequences of the *adh* gene. We chose these datasets in order to complete the model contrasting of previous works [6,10] for the model combinations presented here. These datasets were originally taken from Ref. [20] and we used the same tree topology as therein. Structural descriptors for the SC models were based on the PDB entries 4HHBB and 1A4U for the β-globin and *adh* genes respectively. We used the same priors as described in Ref. [10].

[Rodrigue and Philippe, 2010]

Still to be worked on

- Zaheri, M.; Dib, L. & Salamin, N. A generalized mechanistic codon model Molecular biology and evolution, 2014, 31, 2528-41
- Rodrigue, N. On the statistical interpretation of site-specific variables in phylogeny-based substitution models. Genetics, 2013, 193, 557-64
- Lartillot, N. & Philippe, H. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Molecular biology and evolution, 2004, 21, 1095-109
- Goldman, N., and Z. Yang, 1994 A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol. Biol. Evol. 11: 725736.

Aldous, D. (1985).

Exchangeability and related topics, pages École d'Été de Probabilités de Saint–Flour XIII–1983. Springer, Berlin.

Antoniak, C. E. (1974).

Mixtures of dirichlet processes with applications to bayesian nonparametric problems. *Annals of Statistics*, 2(6):1152–1174.

Blackwell, D. and MacQueen, J. B. (1973).

Ferguson distributions via polya urn schemes. *The Annals of Statistics*, 1(2):353–355.

Blei, D. M. and Jordan, M. I. (2005).

Variational inference for dirichlet process mixtures. *Bayesian Analysis*, 1:121–144.

Escobar, M. D. and West, M. (1995).

Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90(430):577–588.

Ferguson, T. S. (1973).

A bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1(2):209–230.

Kleinman, C. L., Rodrigue, N., Bonnard, C., Philippe, H., and Lartillot, N. (2006).

A maximum likelihood framework for protein design. *BMC bioinformatics*, 7:326.

Lartillot, N. and Philippe, H. (2004).

A bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Molecular biology and evolution*, 21(6):1095–109.

MacEachern, S. N. (1994).

Estimating normal means with a conjugate style dirichlet process prior.
Communications in Statistics B, 23:727–741.

Muse, S. V. and Gaut, B. S. (1994).

A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitutions, with applications to chloroplast genome.
Mol. Biol. Evol, 11:715–724.

Rasmussen, C. E. (2000).

The infinite gaussian mixture model.
In *Advances in Neural Information Processing System*, volume 12.

Robinson, D. M., Jones, D. T., Kishino, H., Goldman, N., and Thorne, J. L. (2003).

Protein evolution with dependence among codons due to tertiary structure.
Mol. Biol. Evol., 20(10):1692–1704.

Rodrigue, N., Kleinman, C. L., Philippe, H., and Lartillot, N. (2009).

Computational methods for evaluating phylogenetic models of coding sequence evolution with dependence between codons.
Molecular biology and evolution, 26(7):1663–76.

Rodrigue, N. and Philippe, H. (2010).

Mechanistic revisions of phenomenological modeling strategies in molecular evolution.
Trends in genetics, 26(6):248–52.

Rodrigue, N., Philippe, H., and Lartillot, N. (2010).

Mutation-selection models of coding sequence evolution with site-heterogeneous amino acid fitness profiles.
Proceedings of the National Academy of Sciences of the United States of America, 107(10):4629–34.

Sethuraman, J. (1994).

A constructive definition of dirichlet priors.
Statistica Sinica, 4:639–650.

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2006).

Hierarchical dirichlet processes.
Journal of the American Statistical Association, 101(476):1566–1581.

Yang, Z. and Nielsen, R. (2008).

Mutation-selection models of codon substitution and their use to estimate selective strengths on codon usage.
Mol. Biol. Evol., 25:568–579.