Notes

Internal Repeat Variability in Mucin Sequences

F. N. Braun*,† and G. C. Hansson‡

Institute of Medical Biology, University of Tromso, N-9037 Tromso, Norway, and Department of Medical Biochemistry, Göteborg University, SE-40530 Gothenburg, Sweden

Received February 8, 2006
Revised Manuscript Received June 1, 2006

The outstanding structural feature of mucins is their heavy glycosylation via O-linkage to repetitive domains of the apoprotein backbone.[1] Glycans constitute 50−80% of the assembled molecule by weight, largely determining the physical chemical properties of mucus, and they are also implicated in immunological function as binding sites for microbial pathogens. Hence, it is of interest to explore the nature of the evolutionary selection pressures which have shaped mucin glycosylation.

A noteworthy feature established recently by Julenius et al.[2] is that the O-linkable amino-acid residues Ser and Thr, which provide attachment sites for the glycans, do not appear to be conserved in any site-specific fashion; that is, selection for such residues must presumably act instead on some more coarse-grained bulk level. Here, toward further elucidation of this issue, we look for evidence of Ser/Thr selection at the coarse-grained level of the repeats. Our ansatz is that selection pressure of this type, if present, manifests itself via the normalized frequency profile ϕ_γ, where γ is the varying total number of Ser/Thr sites per repeat.

In Figure 1a, we plot ϕ_γ for a large 73-repeat domain which dominates the sequence architecture of the human mucin MUC5B.[3] Out of a total $R = 29$ sites in the repeat unit,[4] typically around half are glycosylisable, $\langle\gamma\rangle \approx 15$. Figure 1b interprets this observed profile according to a bare Markovian site-mutational approach, i.e., let $p(\tau)$ and $q(\tau)$ denote respectively the probability that an O-linkable residue flips to any of the 18 remaining amino acids during evolutionary time τ, and vice versa. The long-time steady state is then just the binomial distribution defined by R and the ratio $q^* = p(\tau)/q(\tau)$, with the latter in effect constituting a "hidden" parameter to be determined by some form of tuning to the observed profile. A convenient tuning criterion is just to match the expected versus observed mean; that is, $\langle\gamma\rangle^{(e)} = \langle\gamma\rangle^{(o)}$. Then we have straightforwardly $q^* = \langle\gamma\rangle/(R - \langle\gamma\rangle)$, with expected variance $\sigma^2 = \langle\gamma\rangle - \langle\gamma\rangle^2/R$.

The value $q^* = 1.04$ determined thus for MUC5B indicates that Ser/Thr substitutions occur at a rate which is comparable to all other rates of substitution combined. Since this is obviously far in excess of neutral rates as predicted for example by the well-known Jukes-Cantor scheme,[5] it can be taken as evidence for their "bulk positive selection", independently of site-specific location within the repeat unit. There is room for improvement to this bare description, however, insofar as it gives a poor account of skewness of the observed profile (Table 1). This motivates an extended approach, Figure 1c, where we improve the fit by adding a negative (purifying) mode of selective pressure,[6] acting to truncate the distribution from the right at some cutoff γ_c (see the Methods section).

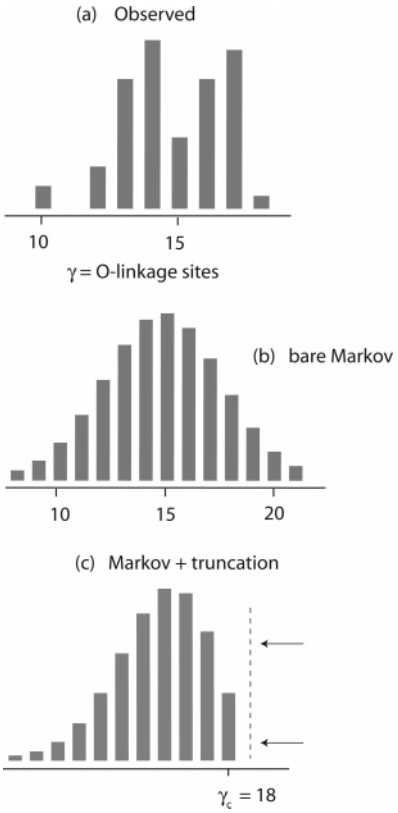

Figure 1. Model interpretation of mucin internal repeat variability with respect to number of glycan attachment sites: (a) observed Ser/Thr profile ϕ_γ for MUC5B; (b) bare Markov (binomial) model fit; (c) improved model fit, introducing a truncation cutoff. The choice $\gamma_c = 18$ corresponds to the highest single instance of γ observed in the MUC5B sequence.

What is the physical−chemical rationale for such a cutoff? If the primary functional role of the apoprotein is to provide a scaffolding for glycans, then what stops it from carrying a much higher density of potential glycosylation sites? A possible explanation concerns the maintenance of a functional viscoelastic texture in mucus, which is known to depend sensitively on the degree of conformational flexibility of the assembled mucins.[7] Mucins tend to increase in rigidity as more glycans are attached; hence, it is feasible to suppose that steric hindrance and the need for conformational flexibility impose an upper limit on the number of attachment sites. This should be qualified, however, by noting that typically not all Ser/Thr sites are occupied in the assembled mucin. The extent of glycosylation appears thus also subject to more subtle affinities of the individual polypeptide−GalNAc transferases responsible for attachment. Twenty such transferases are present in the human genome. Although not yet studied in comprehensive detail, there are indications that at least some of these may require an already

* To whom correspondence should be addressed.
† University of Tromso.
‡ Göteborg University.

10.1021/bm0601169 CCC: $33.50
Published on Web 10/31/2006

Table 1. Quantitative Statistics for Figure 1[a]

	bare Markov			Markov + truncation ($\gamma_c = 18$)	
fit quantity $X[\phi_\gamma]$	observed X^o	expected X^e	goodness of fit $(X^e - X^o)^2/\lvert X^e\rvert$	expected X^e	goodness of fit $(X^e - X^o)^2/\lvert X^e\rvert$
		(q^*=1.04)		(q^*=1.36)	
mean, $\langle\gamma\rangle$	14.77	14.77	0	14.77	0
variance, σ^2	3.25	7.25	2.21	4.08	0.17
skew, μ_3/σ^3	−0.34	−0.01	10.89	−0.52	0.06

[a] In the definition of skew, μ_3 refers to the third central moment of ϕ_γ.

GalNAc-primed peptide. OGLYCBASE, a database of known O-glycosylation sites,[8] suggests a role for specific flanking residues, notably the non-glycosylisable residue proline.[2] Selection for a small number of proline sites in the MUC5B repeat unit can of course be viewed as tantamount to a conjugate truncation of Ser/Thr sites, along the lines we are proposing.

In order to investigate the truncation constraint more generally beyond the specific case of MUC5B, we applied the approach to other members of the human mucin family.[9] A similarly plausible fit is obtained in particular for the sequence of MUC17,[10] a mucin expressed in the intestines (60 tandem repeats of length $R = 61$). Moreover this best fit corresponds to a value $q^* = 1.37$ of the Markov parameter. Since this is very close to that found for MUC5B, it suggests that point-mutational turnover in O-linkage sites remains more or less uniform among mucin paralogs. On the other hand, further inspection of the human mucin family reveals several cases (e.g., MUC1, MUC3A, and MUC4) where $\phi_\gamma^{(o)}$ has a standard deviation of less than one residue. The likely reason for this is a frequent rate of recombinatory events occurring internally within the repetitive domain, and these are not accounted for in our model scheme. As is well-known, if such events occur at a rate much faster than that of point mutation, then they exert a "concerting" influence which tends to wash out point-mutational variability between repeats.[11] We cannot directly test the validity of our model in such cases, since from a purely statistical perspective a healthy degree of variance is requisite for quantitative analysis. MUC5B and MUC17 serve in this respect as a proof of principle, evidence for the existence of selective pressures which in many other mucins may be obscured by concerted evolutionary homogenization.

Methods

In the Markov + truncation model the expected profile, $\phi_\gamma^{(e)}$ is identified as the meaningful eigenstate of the mutational matrix $M_{\gamma\gamma'}(\tau)$ whose elements give the probability that a repeat with composition γ' mutates to composition γ

$$M_{\gamma\gamma'}(\tau) = [1 - \gamma' p(\tau) - (R - \gamma')q(\tau)]\delta_{\gamma\gamma'} + \gamma' p(\tau)\delta_{\gamma,\gamma'-1} + (R - \gamma')q(\tau)\delta_{\gamma,\gamma'+1}$$

where δ_{ij} is the Kronecker delta function, and a negative selective cutoff γ_c limits the index range of the matrix, i.e., $\gamma,\gamma' = 0..\gamma_c$ (note that with increasing γ_c the eigenstate approaches the binomial distribution of the bare model). It is useful to exploit the property that $M_{\gamma\gamma'}$ has the same eigenstates as a simpler matrix $M^*_{\gamma\gamma'}$ (q^*)

$$M^*_{\gamma\gamma'} = \gamma'(q^* - 1)\delta_{\gamma\gamma'} + \gamma'\delta_{\gamma,\gamma'} - 1 + (R - \gamma')q^*\,\delta_{\gamma,\gamma'+1}$$

For given q^*, the eigenstate can be calculated numerically, e.g., by tridiagonal inverse iteration,[12] taking the observed profile as an initial guess.

References and Notes

(1) Carlstedt, I.; et al. *J. Biol. Chem.* **1993**, *268*, 18771−18781.
(2) Julenius, K.; Molgaard, A.; Gupta, R.; Brunak, S. *Glycobiology* **2004**, *15*, 153−164.
(3) Desseyn, J. L.; Aubert, J. P.; Porchet, N.; Laine, A. *Mol. Biol. Evol.* **2000**, *17*, 1175−1184.
(4) MUC5B repeats exhibt length variance $\sigma^2 = 2.4$, which is neglected here (44 of 73 in total have the $R = 29$ consensus length specified)
(5) Lio, P.; Goldman, N. *Genome Res.* **1998**, *8*, 1233−44.
(6) It might be objected that O-linkable sites cannot be subject to both negative and positive selection pressures simultaneously; however, in our view, there is no inconsistency. The extended model Figure 1c distinguishes two separable factors and decouples them. On the one hand, there is the positive selective value of a single O-linkable residue taken in isolation, reflected in a high value of q^*. This is distinct from the overall density of such sites, which is a collectively determined trait subject in principle to a different set of selective constraints.
(7) Bansil, R.; Stanley, E.; Lamont, J. T. *Ann. Rev. Physiol.* **1995**, *57*, 635−657.
(8) Gupta, R.; et al. *Nucleic Acids Res.* **1999,** *27*, 370−372 (www.cbs.dtu.dk/databases/OGLYCBASE).
(9) A curated database of sequences for the mucin genes of the human genome can be found online at http://www.medkem.gu.se/mucinbiology/databases/index.html.
(10) Gum, J. R.; Crawley, S. C.; Hicks, J. W.; Szymkowski, D. E.; Kim, Y. S. *Biochem. Biophys. Res. Comm.* **2002**, *291*, 466−475.
(11) Smith, G. P. *Science* **1976**, *191*, 528−535.
(12) Press, W. H.; et al. *Numerical Recipes in C*; Cambridge University Press: New York, 1992.

BM0601169