

# Molecular Evolution and Network-Level Analysis of the N-Glycosylation Metabolic Pathway Across Primates

Ludovica Montanucci,<sup>1</sup> Hafid Laayouni,<sup>1,2</sup> Giovanni Marco Dall'Olio,<sup>1</sup> and Jaume Bertranpetit<sup>\*,1,2</sup>

<sup>1</sup>Institute of Evolutionary Biology (Universitat Pompeu Fabra-Consejo Superior de Investigaciones Científicas), Department of Experimental and Health Sciences-Universitat Pompeu Fabra-Parc de Recerca Biomèdica de Barcelona, Barcelona, Catalonia, Spain

<sup>2</sup>Centro de Investigación Biomédica en Red de Epidemiología y Salud Pública, Spain

\*Corresponding author: E-mail: jaume.bertranpetit@upf.edu.

## Abstract

N-glycosylation is one of the most important forms of protein modification, serving key biological functions in multicellular organisms. N-glycans at the cell surface mediate the interaction between cells and the surrounding matrix and may act as pathogen receptors, making the genes responsible for their synthesis good candidates to show signatures of adaptation to different pathogen environments. Here, we study the forces that shaped the evolution of the genes involved in the synthesis of the N-glycans during the divergence of primates within the framework of their functional network. We have found that, despite their function of producing glycan repertoires capable of evading rapidly evolving pathogens, genes involved in the synthesis of the glycans are highly conserved, and no signals of positive selection have been detected within the time of divergence of primates. This suggests strong functional constraints as the main force driving their evolution. We studied the strength of the purifying selection acting on the genes in relation to the network structure considering the position of each gene along the pathway, its connectivity, and the rates of evolution in neighboring genes. We found a strong and highly significant negative correlation between the strength of purifying selection and the connectivity of each gene, indicating that genes encoding for highly connected enzymes evolve slower and thus are subject to stronger selective constraints. This result confirms that network topology does shape the evolution of the genes and that the connectivity within metabolic pathways and networks plays a major role in constraining evolutionary rates.

**Key words:** molecular evolution, N-glycosylation, network analysis, degree centrality.

## Introduction

A fundamental issue in molecular evolution is the understanding of the principles that drive the evolution of genes whose products interact in complex functional networks. Indeed, a major limitation of the existing approaches in studying molecular evolution has been the approximation that considers each gene as a single entity and describes the action of natural selection on it without integrating the information about its interaction network. Because biological function is the result of a large number of interacting molecules organized in complex networks and arises as an emergent property from a combined effect of many different genes, it is now widely accepted that steps forward in understanding the complex basis of adaptation can be achieved by integrating the knowledge derived from evolutionary studies into a network framework (Cork and Purugganan 2004). In recent years, the development of network analysis in biology has received increasing interest because many biological systems can be represented as networks of interacting components (for a recent review, see Yamada and Bork 2009).

The detection of the action of selective pressures on individual protein-coding genes can be achieved through the estimation of synonymous and nonsynonymous substitution rates ( $dS$  and  $dN$ , respectively). The ratio  $\omega$ , or  $dN/dS$ , gives the strength and the direction of the action of selec-

tive pressures, indicating accelerated evolution or positive selection, neutral evolution, or negative selection when it is higher, equal, or lower than 1, respectively. This measure can be estimated for the entire coding sequence of a gene as a whole or multiple estimates may be made for different site classes (Yang 2006). Its departure from 1 reflects the strength of the selective pressure and, in absence of events of positive selection,  $\omega$  measures the strength of the purifying selection in preventing the fixation of deleterious replacements, and it is known to correlate with protein dispensability (Hirsh and Fraser 2001) even if it does correlate with a plethora of variables and mainly with expression levels (Rocha 2006). Some studies have investigated the relationship between evolutionary rates and network properties, in order to unravel how purifying selection (or adaptive variation) is distributed over molecular networks. At a large scale, whole metabolic and protein-protein interaction networks have been investigated in relation to the evolutionary rates of their genes. In protein-protein interaction networks, central proteins and highly connected proteins (hubs) have been found to experience high evolutionary constraints (Jeong et al. 2001; Fraser et al. 2002; Hahn and Kern 2005), although positive selection seems to occur at the network periphery (Kim et al. 2007). Furthermore, interacting proteins seem to show similar level of divergence (Fraser et al. 2002; Lemos et al. 2005). In the case of metabolic networks, it has been

found that enzymes that are highly connected and those located at branch points exhibit slower rates of amino acid replacement (Vitkup et al. 2006; Greenberg et al. 2008).

Other studies focused on specific known metabolic or signaling pathways. In an early work, Rausher et al. (1999) investigated the evolutionary rates of the genes of a metabolic pathway, inspecting whether the differences in their rates of evolution could be explained by the properties of the network in which they participate. They found an increase of the rates of evolution along the pathway from upstream to downstream. The slow rates of evolution of the upstream genes were attributed to their greater pleiotropy because, in a linear pathway, they are more likely to be above branch points, thus influencing a higher number of products. Other studies confirm the same pattern of slower rates in the upstream region (Lu and Rausher 2003; Riley et al. 2003; Livingstone and Anderson 2009; Ramsay et al. 2009). This rule also holds for the six genes involved in the Ras-mediated signal transduction pathway, for which the levels of polymorphism in *Drosophila melanogaster* and the levels of divergence with *D. simulans* (Riley et al. 2003) were studied. Ramsay et al. (2009) introduced a new measure of pathway position, the Pathway Pleiotropy Index, which counts groups of enzymes between branch points, which better correlates with evolutionary rate in respect to the simple pathway position. On the contrary, an opposite gradient in the strength of purifying selection has been found by Alvarez-Ponce et al. (2009). The authors examined the insulin/TOR signal transduction pathway across 12 *Drosophila* species with the aim of understanding the impact of network topology on the sequence evolution. They still found strong evidence that the level of functional constraint does depend on the position of each gene in the pathway, but downstream genes appear more constrained than upstream ones. Flowers et al. (2007), studying the evolution of enzymes of five metabolic pathways, observed that those located at branch points are more likely to be target of adaptive selection. Finally, a similar analysis of the gibberellins pathway (Yang et al. 2009) revealed that genes located at major branch points had the lowest evolutionary rates regardless of their upstream/downstream position. No previous studies exist in humans, primates or even mammals.

All these studies support the idea that evolutionary pressures acting on genes are in close relation with the structure of their functional network; however, a general principle could not be drawn because different patterns have been found for different pathways and different species sets. New pathways still have to be analyzed and compared in order to infer local or pathway-specific as well as general rules, to unravel underlying principles and to disentangle topological restrictions of the network from the biological properties and functions.

In this work, we study the N-glycosylation pathway. Glycosylation is one of the most common forms of protein modification and consists of the attachment of a glycan to a protein. Glycans are oligosaccharide chains of sugars that can be attached to proteins or lipids and carry out

important roles in mediating many biological functions of the cell. In nature, glycans are characterized by their complex branched forms, which give rise to an extraordinary structural diversity. In particular, the cell surface is endowed with a dense covering of exposed glycans. Those extracellular glycans mediate both cell–cell and host–pathogens interactions as they can be recognized by viral and bacterial pathogens and parasites that can initiate infection by binding to the glycan surface of host cells. Those recognized by pathogens may be targets of strong selective pressures, giving rise to rapid changes in the repertoire of glycans and the combinatorial nature of their synthesis allows the possibility of making rapid changes to escape the binding of pathogens (Gagneaux and Varki 1999). It has been proposed that rapid evolution of glycans due to the pressure of infectious disease could also contribute to speciation processes (Varki 2006).

Here, we study the genes involved in the synthesis of the N-glycans, which are those attached to the asparagine (N) residue of proteins and account for a large proportion of the glycan variety (Varki et al. 2009). Their synthesis is carried out by a limited number of enzymes in an assembly line–like process (Varki 2006). All N-glycans share a common core sugar sequence, which is named the “precursor”: It consists of 14 sugars (two core GlcNAcs, nine manoses, and three terminal GlcNAcs) and is also referred to as a dolichol lipid–linked oligosaccharide because the dolichol serves as a lipid carrier and as a membrane anchor for the assembly of this oligosaccharide. The first step in N-linked glycosylation is therefore the synthesis of the precursor: It consists of a series of 14 subsequent sugar additions to a phosphorylated phospholipid. The precursor biosynthesis starts at the cytoplasmic side of the endoplasmic reticulum (ER) membrane and then the sugar is flipped across the ER membrane, moving it from the cytosolic side into the ER lumen; this translocation is operated by the flippase protein RFT1. Once the synthesis of the precursor is completed, it is transferred to a nascent protein chain. The N-glycan precursor is attached in a single step to the consensus sequence Asn-X-Thr/Ser of the nascent protein, releasing the dolichyl phosphate anchor and the unfolded glycoprotein. This reaction occurs cotranslationally, and it is catalyzed by the oligosaccharyltransferase (OST), which is a protein complex consisting of several subunits. Following this, there is a “quality control” step, in which the glucosidase II complex and the MOGS ensure that the glycoprotein has properly folded. Finally, the N-glycan precursor, which is now covalently linked to the folded glycoprotein, is extended through sequential attachment of monosaccharides (the glycan extension step), allowing the formation of the enormous variety of different final branched structures. This process is catalyzed by a limited number of enzymes, many of which can accept different glycans as substrates. This last step has been investigated from a network perspective in a recent paper (Kim et al. 2009). The authors analyze a network made of 638 nodes, each one representing a different glycan and 1,499 edges, which represent enzymatic reactions that transform one

glycan into another; the enzymes responsible for these reactions are those involved in the last step of glycan extension. In that work, the authors unravel the extreme modularity of this network by identifying 21 cohesive modules. All these modules are arranged in a centralized structure: A common upstream module acts as a control tower redistributing the metabolic fluxes on the different peripheral ones. However, the more common network analysis in which nodes represent genes has not yet been performed for this system.

Evolutionary rate studies have to focus on a defined group of species that may be of very different size. When done on a wide group, which comprises also distantly related organisms, there may have been main changes of function (like the acquisition of a new function by a protein domain), causing the purifying selection measure to lack biological sense; on the other hand, many statistical tests, such as those for detecting positive selection, may have reduced statistical power when applied on a very small group of closely related species with low number of substitutions. The present study focuses on a small group of primate species for which there is an advanced whole-genome sequencing project and for whose genes it is possible to assume functional conservation.

A clean and complete data set is essential for the reliability of any study; in particular, evolutionary analyses are extremely sensitive to the goodness of the multiple alignments on which they are based: Even exiguous errors (that may arise either from the sequencing or from the annotation process) can severely affect and dramatically bias the results. Even if data from whole-genome projects have a good quality at large scale, they often harbor minor errors and inaccuracies that turn out to be crucial in driving the observed evolutionary signals. Hence, it is not possible to rely only on automatic procedures, and careful data set cleaning is necessary.

The aim of this work was to analyze the evolutionary constraints of the N-glycan pathway in order to understand its dynamics, dispensability of the components, and biological and topological implications, with an overview on the global evolvability of the pathway.

## Materials and Methods

### Data Set

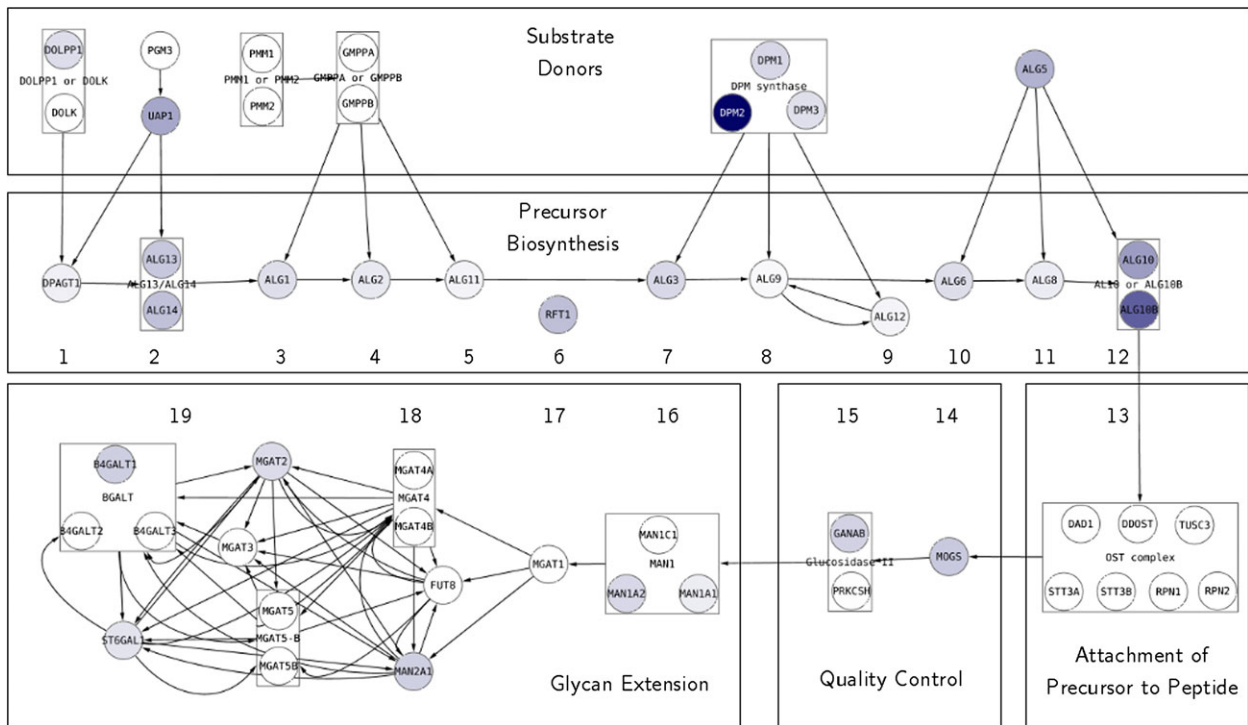
#### Genes

The data set is composed of the genes involved in the N-glycan biosynthesis pathway. The identifiers of the genes involved in this pathway were downloaded from the KEGG database (PATHWAY: hsa00510). Among the 46 genes listed in KEGG, only the *MAN1B1* was excluded from the analysis because its sequences from nonhuman primates contained suspicious regions that yielded a number of mismatches and gaps in the final alignment up to 71% of the alignment length. This gene list was manually integrated with another seven genes, which were not included in the KEGG source but were derived from literature. The data set is therefore composed of the 52 genes listed in

**Table 1.** List of the Genes of the N-Glycosylation Biosynthesis Pathway Used in This Study. The Length of the Encoded Proteins, the Percentage of the Used Codons for the Evolutionary Analysis, and the Codon Bias [effective number of codons (ENC)] Are Also Reported. In the Last Three Columns,  $\omega$  Values,  $dN$ , and  $dS$  for Each Gene Under the Model M0.

	Gene	Length	% Used					
			Codons	ENC	$\omega$	$dN$	$dS$	
Precursor biosynthesis	ALG1	464	81.5	49.28	0.201	0.032	0.16	
	ALG2	416	100	52.74	0.169	0.023	0.136	
	ALG3	438	88.4	42.95	0.222	0.024	0.107	
	ALG6	509	82.5	51.07	0.205	0.018	0.09	
	ALG8	526	85.4	56.31	0.16	0.011	0.069	
	ALG10	473	89.2	52.15	0.408	0.035	0.086	
	ALG10B	473	78.4	52.27	0.603	0.031	0.051	
	ALG11	492	84.3	48.57	0.134	0.007	0.05	
	ALG12	488	71.7	43.14	0.13	0.027	0.207	
	ALG13	1137	67.6	52.1	0.276	0.023	0.083	
	ALG14	216	99.5	57.12	0.293	0.034	0.115	
	ALG9	843	72.2	54.34	0.118	0.009	0.076	
	RFT1	541	99.8	52.9	0.302	0.024	0.078	
	Substrate donors	ALG5	324	100	55.27	0.343	0.025	0.074
DOLK		538	100	46.9	0.062	0.006	0.088	
DOLPP1		238	64.7	46.14	0.207	0.012	0.057	
DPAGT1		408	97.5	49.28	0.151	0.009	0.063	
DPM1		295	72.9	54.63	0.225	0.014	0.064	
DPM2		130	98.5	47.97	0.89	0.039	0.043	
DPM3		122	75.4	36.89	0.198	0.015	0.076	
GMPPA		473	59.4	50.03	0.057	0.005	0.091	
GMPPB		387	93	41.98	0.037	0.004	0.1	
PGM3		570	95.1	52.18	0.089	0.003	0.039	
PMM1		262	84.4	41.53	0.04	0.006	0.141	
PMM2		246	98.4	52.7	0.111	0.013	0.12	
UAP1		522	85.6	54.13	0.376	0.083	0.22	
OST complex		DAD1	113	100	52.76	0	0	0.05
	DDOST	456	66.4	48.77	0.104	0.013	0.121	
	RPN1	607	100	48.44	0.084	0.008	0.101	
	RPN2	631	98.6	50.86	0.096	0.008	0.085	
	STT3A	705	90.9	52.73	0.018	0.001	0.083	
	STT3B	826	87.4	55.25	0.046	0.003	0.069	
	TUSC3	348	85.1	57.73	0	0	0.046	
	Quality control	GANAB	966	100	51.12	0.241	0.017	0.069
		MOGS	837	93.7	51.53	0.268	0.021	0.078
		PRKCSH	535	77.9	43.92	0.105	0.024	0.224
Glycan extension	B4GALT1	398	89.4	49.59	0.252	0.031	0.125	
	B4GALT2	401	80.3	40.93	0.059	0.008	0.135	
	B4GALT3	393	93.9	50.87	0.088	0.005	0.061	
	FUT8	575	65.4	54.28	0.039	0.003	0.067	
	MAN1A1	653	77.2	55.38	0.155	0.014	0.089	
	MAN1A2	641	100	54.05	0.221	0.01	0.047	
	MAN1C1	630	57	42.86	0.063	0.009	0.134	
	MAN2A1	1144	73.1	52.94	0.246	0.021	0.086	
	MGAT1	445	67.9	38.13	0.064	0.02	0.309	
	MGAT2	447	100	52.31	0.213	0.007	0.035	
	MGAT3	533	93.8	31.81	0.029	0.012	0.425	
MGAT4A	535	88.8	52.57	0.031	0.003	0.094		
MGAT4B	563	48.1	38.88	0.021	0.005	0.216		
MGAT5	741	89.3	52.92	0.076	0.005	0.069		
MGAT5B	801	89.9	35.75	0.051	0.015	0.286		
ST6GAL1	406	100	51.86	0.186	0.018	0.099		

table 1. We classified the gene-based components of the pathway into five functional classes according to their role within this process (see fig. 1): substrate donors, comprising genes that provide the monosaccharide sugar to the enzymes that will catalyze the reaction of its addition to



**Fig. 1** Representation of the N-glycan biosynthesis pathway. Each node represents a gene and each edge a metabolic reaction. Each gene is colored in a gradient according to its levels of  $\omega$  (white for  $\omega < 0.1$  and dark blue for  $\omega > 0.9$ ). Intermediated values are shown in a gradient between the two colors. Numbers indicate subsequent steps in the biosynthesis of the glycan.

the glycan precursor; precursor biosynthesis, comprising genes encoding for the enzymes responsible for the synthesis of the precursor; the OST complex, which catalyzes the attachment of the glycan precursor to a nascent polypeptide chain; quality control, made up of three genes that ensure the proper folding of the nascent glycoprotein; and finally, glycan extension, constituted by the genes that are involved in the extension of the glycan precursor.

### Orthologies

For each one of the 52 human genes, we sought its orthologs from four other primate species: *Pan troglodytes* (chimpanzee), *Gorilla gorilla* (gorilla), *Pongo pygmaeus* (orangutan), and *Macaca mulatta* (macaque) in the Ensembl-Compara database (release 57). We filtered genes having a 1:1 orthology relationship with the human reference gene. An exception was made for the *ALG1* gene, for which multiple orthologous genes were reported in the orangutan and macaque genomes: Because no full sequence gene duplication seemed to have occurred (Marquès-Bonet T, personal communication), it was possible to identify and retain the sequences of the functional genes. The macaque orthologs of *ALG10* and *ALG10B* genes are annotated as 1: many since the *ALG10* gene underwent a duplication specific to the great apes (it is duplicated in human, chimpanzee, gorilla, orangutan but not in macaque). The same macaque sequence has been therefore kept for both *ALG10* and *ALG10B*. For three genes, we could not retrieve the ortholog for all the four nonhuman primate species:

*MGAT2* lacks an annotated ortholog for both gorilla and macaque, whereas *ALG11* and *MOGS* lack an annotated ortholog for gorilla. However, these are likely to be annotation problems rather than true cases of gene loss/gain because both the gorilla and the macaque genome projects do not reach the high annotation quality of the human, the chimpanzee, and the orangutan ones. Although searches against the whole genome were attempted in order to predict them, no homologous sequence could be retrieved by similarity search.

### Sequences

For each gene, we focused the analysis on the protein-coding DNA sequence (CDS) of its longest transcript, retrieved from Ensembl (release 57). Given the low divergence among the considered species, the length of the sequences and the sequences themselves are highly conserved. However, some sequences from nonhuman primates show long deletions (up to 300 bp) in the middle of the gene when aligned to their orthologous group. We suspected that many of these cases, instead of being true cases of deletion, were likely to be artifacts due to a low quality of the sequence or of the annotation that can lead to an artificial absence of a sequence region or of a whole exon. When a CDS was showing absence of one or more exons and therefore was suspected to be incomplete, we tried to recover the putative missing parts of the sequence through a similarity search-based procedure: first, a BLAT search against the whole genome of interest on the UCSC Genome Browser was performed

(<http://genome.ucsc.edu/cgi-bin/hgBlat>); if not successful, a Blast search against NCBI Traces of whole genomes was performed (<http://blast.ncbi.nlm.nih.gov/Blast.cgi>). When an homologous genomic region could be retrieved, the structure of the gene was predicted with the Wise2 program of the GeneWise tool (Birney et al. 2004) applied with default options. This program allows prediction of the structure of a gene (introns–exons) given a genomic DNA region and a protein sequence of high homology to the putative one. The genomic DNA region identified through the BLAT search and the protein sequence of the human reference gene were given as input to Wise2, which predicted the gene structure. Only good predictions in which there were no internal stop codons or frame shifts were accepted. Gene order around the target gene was investigated to corroborate this prediction. With this procedure, we were able to retrieve the missing part of the CDS for three genes: the *DPM2* orangutan gene was missing 42% of the sequence in respect to the human gene, the *ALG13* macaque gene was missing 62%, and the *ALG3* was missing 48%.

### Multiple Alignments and Quality Assessment

The multiple alignments of the CDS of each orthologous sequences were downloaded from the Ensembl-Compara database (release 57) as computed by the GeneTrees method (Vilella et al. 2009) through the Perl API. In the cases where the sequences were improved through prediction, the multiple alignment of the sequences of each orthologous set was obtained with the T-coffee program (Notredame et al. 2000). This has been the case of three genes: *DPM2*, *ALG3*, and *ALG13*. The CDS sequences were aligned as protein sequences (t\_coffee command with default options) and then back translated to the original DNA sequences.

Because the quality of the alignment dramatically affects the estimation of the evolutionary parameters, and because spurious sequence stretches are likely to give rise to false signals of positive selection, we decided to restrict the evolutionary analysis to well-aligned regions. The cleaning procedure was carried out through the Gblocks program (Castresana 2000), which recognizes and removes poorly aligned regions by identifying contiguous positions in the alignment with a suspiciously high number of differences (the default is half the number of sequences plus one). However, this program, when applied to a multiple alignment, is not able to recognize when a single sequence shows an extensive region (contiguous positions) with an unusually high divergence (e.g., as we often found for the macaque sequences). In order to automatically detect such regions through Gblocks, we applied the software to the pairwise alignments of the human sequence and each one of its orthologs from the other four species. By considering pairwise alignments, substrings with low divergence in a single sequence can be detected and masked. All the positions of the ortholog sequence that have been masked by the program were also masked in the original multiple alignment and therefore those regions were not

considered in the subsequent evolutionary analysis. Gblocks was applied with the following parameters:  $-b3 = 3$ ,  $-b4 = 20$ . Gblocks removed bad aligned regions from 34 (of a final data set of 52) alignments; for 30 of these, the percentage of masked sites was below 10% of the alignment length, the maximum percentage of masked sites being 26% for the *ALG12* gene. The information loss due to the masking procedure is therefore not considerable and the reliability of the multiple alignments data set is strongly improved. In table 1, the percentage of the sequence used for the evolutionary analysis is reported.

### Graph Representation of the Pathway

The structure of this metabolic pathway was derived from the KEGG pathway database and was manually extended through the integration of interactions derived from literature; it is represented in Figure 1, which was produced with Cytoscape version 2.6.3 (Shannon et al. 2003). Nodes represent gene products and are indicated with their gene names. When gene products carry out their function within protein complexes, they appear surrounded by a black box. Three protein complexes are present in this pathway: the DPM synthase, formed by the three subunit encoded by *DPM1*, *DPM2*, and *DPM3*; the glucosidase II formed by the catalytic subunit *GANAB* and the regulatory subunit *PRKCSH* and finally the OST complex composed of gene products from *DAD*, *DDOST*, *RPN1*, *RPN2*, *TUSC3*, and *STT3A* or *STT3B* genes. The other boxes surround groups of genes that can alternatively carry out a function.

Two nodes are connected by an edge if the two corresponding genes share the same metabolite either as substrate or as product (Vitkup et al. 2006). The edges have been derived from KEGG or from literature for the added genes. The edges among the genes belonging to the glycan extension class have been derived from the paper by Kim et al. (2009). *RFT1* is the only gene, which has no links because it does not catalyze any metabolic reaction, instead it mediates the translocation of the sugar from the cytosolic side of the membrane to the ER lumen after the two mannose additions catalyzed by *ALG3* and *ALG11*.

### Analysis of the Evolutionary Rates

The action of natural selection was investigated through the estimation of  $dN$ ,  $dS$ , and their ratio  $\omega$  with the codeml program of PAML package version 4.2 (Yang 2007), which performs a maximum likelihood computation of these evolutionary parameters. The  $\omega$  values can be interpreted as measures of the strength of the purifying selection if the occurrence of positive selection (as footprint of molecular adaptation) has been ruled out. Thus, two analyses were carried out: detecting the existence of positive selection and measuring purifying selection.

For each gene, two tests of positive selection were performed. The first is the most conservative among those provided in the package: It compares models *M1a* (nearly

neutral) and M2a (positive selection). Model M1a postulates two different classes of sites evolving at a different  $\omega$ , the first being  $\omega_0 = 1$  for neutrally evolving sites and the second  $0 < \omega_1 < 1$  for constrained sites. Model M2a adds another class of sites whose  $\omega$  is free to be greater than 1 ( $\omega_2 > 1$ ). The two models are compared through a likelihood ratio test (LRT) to test whether the alternative hypothesis of positive selection explains the data significantly better than the null hypothesis of neutral evolution. A second test of positive selection compares models M7 (null model of neutral evolution), which assumes that  $\omega$  follows a (discrete) beta distribution among sites and M8 (selection model), which adds a class of  $\omega$  which can be greater than 1. Like before, the two hypotheses are compared through a LRT. The false discovery rate correction for the multiple tests of positive selection was achieved through the  $q$  value program of the R package (Storey and Tibshirani 2003). All the computations were repeated five times to assess whether the results were stable. In all the analyses, the F3X4 codon frequency model was used. Three different  $\omega$  (0.1, 1, and 2) were used as starting point for the computation. A unique  $\omega$  encompassing all the branches of the tree and for the entire length of the sequence was computed through the M0 model of the codeml program to estimate the strength of purifying selection acting on the gene.

### Network-Level Analysis

To analyze the evolution of each gene within the context of the structure of the network, we computed, for each node, different topological parameters and calculated their correlation with the evolutionary rate estimates given by model M0.

Because the N-glycan synthesis is a highly linear and sequential process based on a series of successive sugar additions, we attributed a position to each step of the sugar additions. We excluded from this analysis the genes of the class of the substrate donors, which are not directly involved in the metabolic reactions that synthesize the glycan because they just provide the sugar substrates. The genes belonging to all other classes were considered and their positions, determined by successive additions of sugars, are indicated by the ordinal numbers (from 1 to 19) showed in Figure 1. Gene products participating in complexes or carrying out the same function at the same stage are given the same ordinal position. In the last part of the network (genes involved in the glycan extension process), the linearity of the process is broken because many of the enzymes of this class can accept as substrates different glycan structures allowing the formation of the same final glycan structure through the action of different sequences of the same set of enzymes (Kim et al. 2009); in this case, these genes are given the same ordinal position.

For each node (corresponding to a gene), three different measures of centrality were computed: degree centrality (fraction of nodes it is connected to), betweenness centrality (fraction of all shortest paths that pass through that node), and closeness centrality (reciprocal of the average

distance to all other nodes). These measures are descriptions of the topological importance of a node in a graph, given its structure. These computations were achieved by means of the NetworkX package of the Python programming language ([networkx.lanl.gov/index.html](http://networkx.lanl.gov/index.html)).

We also investigated whether genes that are neighbors in the network have related values of  $\omega$ ,  $dN$ , or  $dS$ . To this purpose, we computed the pairwise distances between all the genes defined as the shortest paths between the two corresponding nodes in the graph. These distances were computed through the NetworkX package and collected into a symmetric matrix. Gene products involved within the same complex were given distance 0. We compared the matrix of the gene distances with similar matrices of absolute pairwise gene differences in  $\omega$ ,  $dN$ , or  $dS$  through a standardized Mantel test (Sokal and Rohlf, 1995). The Mantel test computation was performed with the ecodist library within the R package (Goslee and Urban 2007) by randomly permuting 99,999 times the rows and columns of one of the two matrices being compared.

### Multivariate Analysis

To evaluate the importance of relationship between evolutionary estimates and the different descriptive network properties, we performed a multivariate analysis, first through partial correlations and then using the path analysis method implemented in Amos 6.0 software to further investigate these correlations under different causal models. Path analysis is an extension of multiple regression analysis that allows decomposing the regression coefficients into their direct and indirect components by considering an underlying user-defined causal model. This allows the assessment of the statistical significance of the relevant direct components. We therefore performed path analysis in order to disentangle which is the main factor influencing the trends in  $dN$  and  $\omega$  because both negatively correlates with pathway position and degree centrality. The causal model adopted also includes codon bias and protein length and is presented in Figure 4. Degree centrality, pathway position, codon bias, and protein length were considered as exogenous variables, whereas  $dN$  and  $\omega$  as endogenous variables. Because path analysis is highly dependent on the chosen causal model, we repeated the analysis considering codon bias as an endogenous variable.

## Results

### Analysis of the Evolutionary Rates

The first step of the analysis was to identify genes that carry sequence evidence of positive selection. The two tests of positive selection (M1a vs. M2a and M7 vs. M8) were always in agreement, thus giving confidence on the robustness of the results. No signal of positive selection was detected in any of the 52 analyzed genes, though DPM2 yielded a marginal significant  $P$  value (0.043), which was no longer significant after correcting for multiple test comparisons.

**Table 2.** Mean of  $\omega$ ,  $dN$ , and  $dS$  in Each Functional Class of Genes with the 95% Confidence Intervals.

Gene group	$\omega$	CI $_{\omega,95\%}$	$dN$	CI $_{dN,95\%}$	$dS$	CI $_{dS,95\%}$
Precursor biosynthesis	0.243	0.089	0.023	0.006	0.102	0.029
Substrates	0.214	0.140	0.018	0.013	0.090	0.029
OST complex	0.050	0.042	0.005	0.005	0.079	0.025
Quality control	0.205	0.217	0.021	0.009	0.124	0.216
Glycan extension	0.112	0.045	0.012	0.004	0.142	0.059

Once the tests of positive selection were performed, we computed the  $\omega$ ,  $dN$ , and  $dS$  for each gene under the M0 model (table 1). The majority of the genes of this pathway are subject to strong purifying selection with an overall mean  $\omega$  of 0.169. Only 6 genes of 52 (12%) have an  $\omega$  value greater than 0.3, with a maximum value of 0.89 for *DPM2* gene, whereas *ALG10B*, *ALG10*, *UAP1*, *ALG5*, and *RFT1* have  $\omega$  values equal to 0.603, 0.408, 0.376, 0.343, and 0.302, respectively. Finally, two genes, *DAD1* and *TUSC3*, have  $\omega$  values equal to 0, meaning that no nonsynonymous substitutions were observed. These two genes are involved in the attachment of the precursor to the peptide chain and are part of the OST complex.

### Analysis of the Classes of Genes

To test whether the measures of  $\omega$ ,  $dN$ , or  $dS$  were statistically different between the five functional classes that we defined above (means and confidence intervals are shown in table 2), a Kruskal–Wallis test was performed (table 3). The test gave no significant differences in  $dS$  among the five classes. However, the comparison between  $dN$  values of different gene classes were significant. The nonsynonymous substitution rate ( $dN$ ), though, was significantly affected by gene class ( $P = 0.002$ ):  $dN$  of the genes of the precursor biosynthesis class appeared to be higher than the genes of the OST complex. This difference remained significant after correcting for multiple testing. When we considered the  $\omega$  estimates, we also found a significant effect ( $P = 0.0017$ ): The precursor biosynthesis class appeared to have a significantly higher mean value of  $\omega$  than both the OST complex genes and the glycan extension ones. However, we noticed that the first class of precursor biosynthesis includes the genes *ALG10* and *ALG10B*, which stem from a recent duplication, an event that is expected to cause a relaxation of selective constraints; fitting with this expectation, both display relatively high values of  $\omega$  (0.408 and 0.603). We thus repeated the test excluding the values for these two genes, and the two Kruskal–Wallis tests for  $\omega$  and  $dN$  still found marginal significant differences between classes ( $P = 0.048$  and  $P = 0.046$  for  $dN$  and  $\omega$ , respectively). In both cases,

**Table 3.** Results for Kruskal–Wallis Test for  $\omega$  and  $dN$ , for the Five Functional Class of Genes.  $P$  Value Is the Significant Level of the Statistical Test for the Pairwise Comparison After Correcting for Multiple Testing By Ranks.

	$P$	Significantly Different Group Pairs
$\omega$	0.0017	Precursor biosynthesis—OST complex Precursor biosynthesis—Glycan extension
$dN$	0.0020	Precursor biosynthesis—OST complex

**Table 4.** Spearman’s Rank Correlation Coefficient ( $\rho$ ) of Different Measures of Centrality (Degree, Betweenness, and Closeness) and the Position along the Pathway with  $\omega$ ,  $dN$ , and  $dS$ .

	Degree		Betweenness		Closeness		Position	
	$\rho$	$P$ Value	$\rho$	$P$ Value	$\rho$	$P$ Value	$\rho$	$P$ Value
$\omega$	-0.484	0.002	0.251	0.123	-0.475	0.002	-0.375	0.017
$dN$	-0.54	0.0004	0.164	0.319	-0.585	<0.0001	-0.342	0.031
$dS$	-0.139	0.399	-0.084	0.613	-0.174	0.289	0.123	0.448

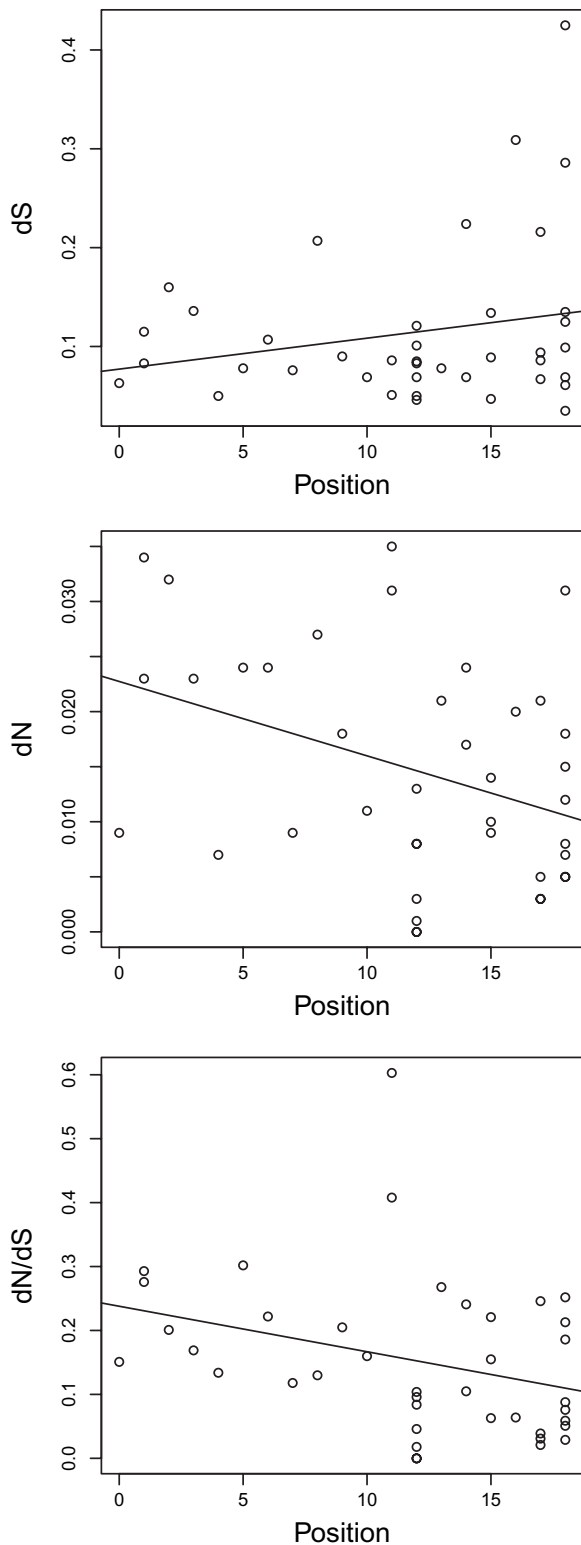
genes of the precursor biosynthesis class appeared to have higher values of  $\omega$  and  $dN$  than those of the OST complex.

### The Strength of the Purifying Selection and Pathway Structure

The estimates of the strength of purifying selection and the synonymous and nonsynonymous substitution rates were analyzed in relation to the structure of the pathway: Each node, representing a gene product, was characterized with different evolutionary parameters ( $dN$ ,  $dS$ , and  $\omega$ ), and we sought to determine whether these parameters correlated with other attributes of the nodes that derive from the structure of the network.

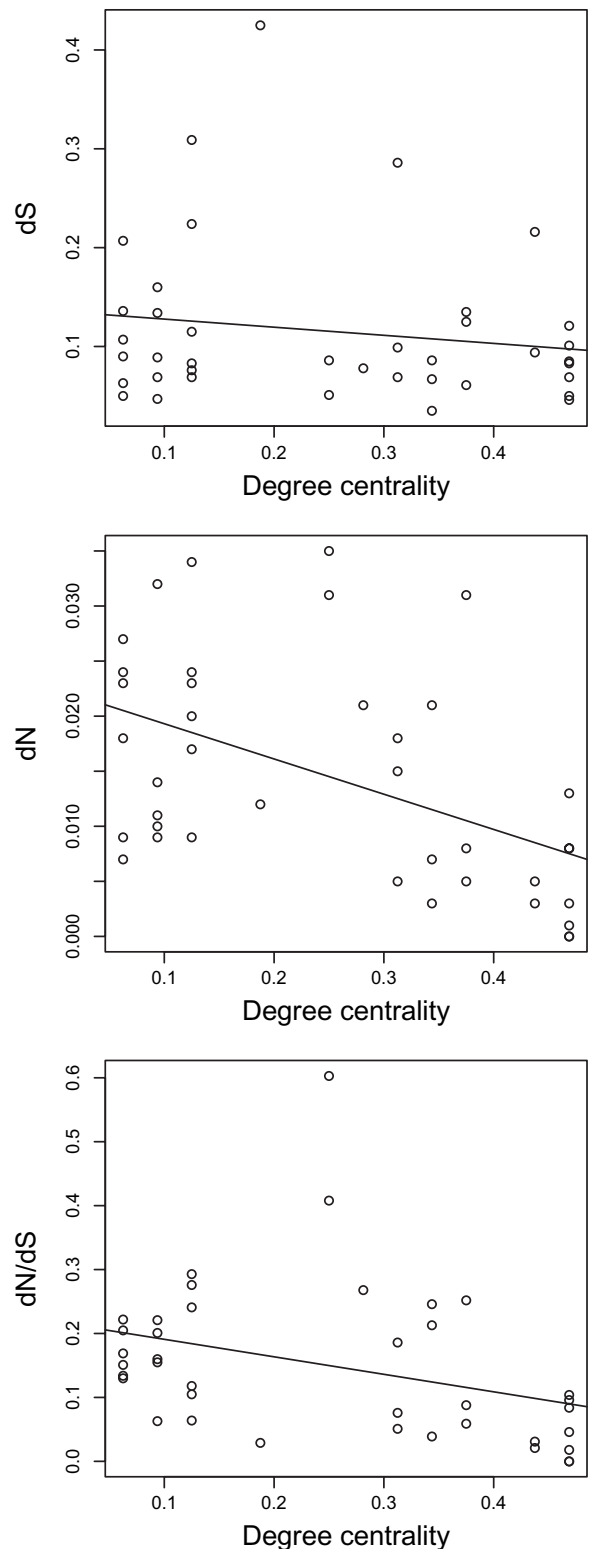
First, because the pathway has a clear sequentiality (given by the successive attachments of the monosaccharide groups in the synthesis of the glycan), we tested whether there was any correlation between the evolutionary measures of a gene and its position along the pathway (from upstream to downstream). Spearman’s rank correlation coefficients between the evolutionary parameters ( $\omega$ ,  $dN$ , and  $dS$ ) and the positions along the pathway (taken by the ordinal number of fig. 1) are reported in table 4 and the plots can be seen in Figure 2. There was a significant negative correlation ( $r = -0.375$ ,  $P = 0.017$ ) between pathway position and  $\omega$ , meaning that upstream genes are subject to relaxed constraints with respect to downstream ones. A negative correlation was also found between  $dN$  and pathway position ( $r = -0.342$ ,  $P = 0.031$ ), whereas no correlation was found between pathway position and  $dS$ . Thus, the strength of purifying selection increases from upstream to downstream genes explicitly due to a decrease in the rate of nonsynonymous substitution.

In a second analysis, different parameters that derive from the topology of the network were computed for the nodes of the network and were investigated in relation to the  $\omega$ ,  $dN$ , and  $dS$  estimates of the corresponding genes (table 4). No significant correlation was found between the betweenness centrality and the evolutionary estimates. However, degree centrality and closeness centrality do correlate with  $dN$  and  $\omega$ . These two centrality measures are expected to correlate with each other because they both depend on the number of edges (degree or connectivity) of each node. In Figure 3 are shown the plots of the evolutionary parameters versus the degree centrality. These results indicate that both  $\omega$  and  $dN$  are negatively correlated with the number of connections of each node, thus the more connected a node is, the lower both its nonsynonymous substitutions rate and its  $\omega$  are.



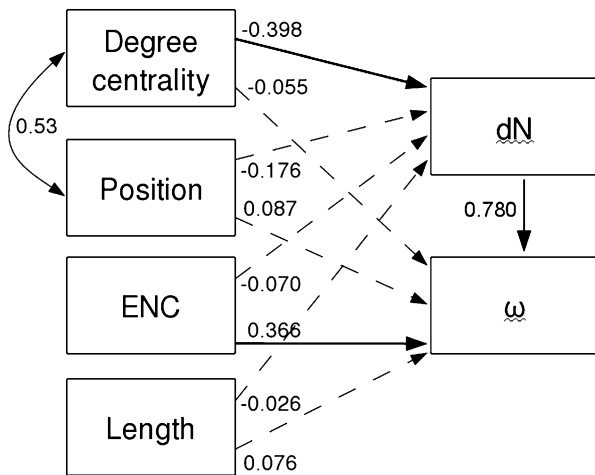
**FIG. 2** The  $\omega$  ( $dN/dS$ ),  $dN$ , and  $dS$  versus pathway position. The correlations are significant for  $\omega$  and  $dN$ . This shows that the strength of the purifying selection increases from upstream to downstream.

A Mantel test comparing a matrix of pairwise distances between genes in the network and matrices of pairwise absolute differences in evolutionary parameters found no significant correlation between distance and difference in  $\omega$



**FIG. 3** The  $\omega$  ( $dN/dS$ ),  $dN$ , and  $dS$  versus degree centrality. The correlations are significant for  $\omega$  and  $dN$  and show that the strength of the purifying selection is higher for more connected genes.

or  $dS$ . However, a significant correlation was found between the distance and difference in  $dN$  ( $r = 0.116$ ,  $P = 0.041$ ). This result supports the idea that neighboring genes share similar evolutionary constraints through the nonsynonymous substitution.



**Fig. 4** Representation of the causal model adopted for the path analysis. Degree centrality, position, codon bias (measured as effective number of codons), and protein length are considered exogenous variables. Numbers on the arrows represent the standardized regression weights. Continuous and dashed lines represent significant and nonsignificant relationships, respectively; single-headed and double-headed arrows represent causal dependencies and correlations, respectively.

### Multivariate Analysis

The analysis of the evolutionary parameters and the pathway structure yielded two main results: significant correlations between the strength of purifying selection ( $\omega$ ) and both pathway position and centrality of genes (calculated as degree centrality). Given the structure of the pathway (fig. 1), it is clear that genes with many connections cluster in the downstream region of the network or, in other words, position and centrality measures correlate (with a Spearman's correlation coefficient  $r = 0.469$  and  $P = 0.003$ ). We therefore sought whether degree centrality and pathway position have by themselves a significant effect on  $\omega$  and  $dN$ , being responsible for constraints on them, or one effect is a side effect of being correlated to the other. Partial correlation revealed that when controlling for position the correlation between degree centrality and  $\omega$  was no longer significant ( $r = -0.242$  and  $P = 0.144$ ) nor was the correlation between position and  $\omega$  when controlling for the degree centrality ( $r = -0.142$  and  $P = 0.396$ ): similarly, when controlling for the degree centrality, the correlation between position and  $dN$  is no longer significant ( $r = -0.129$  and  $P = 0.44$ ). Instead, the correlation between degree centrality and  $dN$  holds significant when controlling for the position ( $r = -0.384$  and  $P = 0.017$ ). This result ensures that the main observed effect is the correlation between degree centrality and  $dN$ , the degree centrality being an important factor affecting the rate of nonsynonymous evolution.

However, other factors could be influencing the rates of evolution. In particular, we investigated whether the relationships between evolutionary rates and topological parameters could be affected by differences in protein length and in codon bias measured as effective number of codons. We found that protein length does not correlate

with any other variable, whereas codon bias is found to negatively correlate with  $dS$  ( $r = -0.626$  and  $P < 0.001$ ). Thus, these two factors do not seem related to the observed correlation of the degree centrality and  $dN$ .

To better characterize the relationships between these variables, we performed a path analysis under the model presented in Figure 4. Similar to what has been found through the analysis of partial correlations, the path analysis revealed that the  $dN$  values are clearly affected by degree centrality ( $P = 0.015$ ), whereas the correlations between position and  $dN$  and between position and  $\omega$  were no longer significant once the effect of degree centrality was removed ( $P = 0.278$  and  $P = 0.407$ ). In this analysis, we also notice that the codon bias is significantly correlated with  $\omega$  ( $P < 0.001$ ), probably acting through  $dS$ . The results are not significantly different under a model where the codon bias is considered an endogenous variable. Among the features that describe the network, the degree centrality is the main factor shaping the nonsynonymous evolutionary rate.

### Discussion

The present evolutionary analysis was performed with the aim of measuring the relative importance or “evolutionary dispensability” of each gene, which indicates to which extent a given protein has accepted amino acid changes through evolution, within a functional network framework. The action of natural selection is at the base of different amounts of genetic dispensability (in the cases of negative or purifying selection) or of adaptation (in cases of positive selection and in the special case of balancing selection). Once the action of positive selection has been ruled out, the strength of purifying selection acting on a gene is given by the amount of conservation in the protein sequence, which can be measured through the  $\omega$  value, which is the ratio of the nonsynonymous ( $dN$ ) versus the synonymous ( $dS$ ) substitution rate. Functional protein-coding genes are expected to be usually under the action of the purifying selection to maintain function, which yields  $\omega$  values much lower than one. However, even with  $\omega$  lower than one, a gene could have experienced positive selection because adaptive forces could have involved relatively few functional sites. In order to reveal the action of selective pressures, LRTs under sophisticated evolutionary models have been developed to test whether an alternative hypothesis of positive selection explains the data significantly better than the null hypothesis of neutral evolution (Yang and Bielawski 2000; Yang et al. 2000; Swanson et al. 2003).

Evolutionary analysis carried out independently on each gene of the whole N-glycan pathway excluded the action of positive or adaptive evolution on these genes during the time of divergence between the Old World primates. Only one gene, the *DPM2*, shows a high  $\omega$  value (0.89); however, no significant signal of positive selection was detected. Two other genes, *ALG10* and *ALG10B*, have  $\omega$  values greater than 0.4. These genes have experienced a recent great

ape-specific duplication, which may be responsible for the relaxation of their selective constraint. It is interesting to note the relative values of  $\omega$  (0.408 and 0.603) for the products of a gene duplication: both genes present a high value, supporting a model in which both copies of a gene after a duplication event start accumulating changes at higher rates (Innan and Kondrashov 2010 and references therein).

The evolution of all genes in the pathway has been mainly driven under purifying selection conditions. Given that the pathway is strongly conserved among the species and that no positive selection has been detected, it is possible to exclude functional shifts and hence through the  $\omega$  value, the strength of purifying selection in preventing the endurance of deleterious variants can be safely estimated. Thus,  $\omega$  is a good measure of evolutionary dispensability of each component (Wilson et al. 1977; Andrés et al. 2007).

Beside the generally high conservation of the pathway, a gene-based analysis has been able to highlight that the class of genes responsible for the extension of the glycan is more constrained than that responsible for the synthesis of the precursor. These genes are responsible for the generation of the great diversity of the N-glycan products, enabling the evasion of rapidly evolving pathogens. Although genes of the glycan extension class would seem to be good candidate targets for positive selection, here we find that they are highly constrained. Instead, gene products that catalyze the synthesis of the precursor are subject to a less strict conservation.

When the strength of purifying selection acting on the genes was analyzed within the context of the network structure of the metabolic pathway, a strong and highly significant correlation was found between the degree centrality of a gene and the number of nonsynonymous changes. Gene products that interact with many others, sharing metabolites with many other gene products, are subject to stronger constraints in their evolution and highly connected enzymes evolve more slowly. This finding had been detected in large-scale studies of the whole yeast (Vitkup et al. 2006) and *Drosophila* (Greenberg et al. 2008) metabolic networks. We also found that genes located downstream in the pathway are subject to a stronger purifying selection than those upstream, showing lower levels of  $\omega$  and of  $dN$ . This result is in agreement with those found for the Insulin/TOR pathway (Alvarez-Ponce et al. 2009) and is contrary to that found for other metabolic and signaling pathways in earlier works (Rausher et al. 1999; Riley et al. 2003; Ramsay et al. 2009). Interestingly, in our case, it has been possible to disentangle the effect of the position along the pathway and connectivity; it is clear that degree centrality (and not the position of the gene along the pathway) is the main causal base of the distribution of  $\omega$  values along the network, stressing the importance of topological and interaction factors in addition to more functional ones.

In this study, we provide more evidence of a tight link between pathway structure and evolutionary forces. Indeed, the evolutionary patterns that we found on these

genes can be a function of the structure of the metabolic network; notably, the degree centrality is a major factor in constraining the evolutionary rate of the genes in this pathway. Whether these findings are specific for this pathway (illuminating the relative importance of its components) and for the species analyzed (primates), or whether they represent a general molecular footprint of functional pathways, remains to be seen in the future, through the accumulation of many more similar analyses to understand molecular evolution within a network.

## Acknowledgments

We thank Pan-Jun Kim and Hawoong Jeong for providing us with information on the glycan extension part of the N-linked glycosylation pathway. We are also grateful to Brandon Invergo for reviewing the manuscript. This research was funded by grants SAF2007-63171 and BFU2010-19443 (subprogram BMC) awarded by Ministerio de Ciencia y Tecnología (Spain) and by the Direcció General de Recerca, Generalitat de Catalunya (Grup de Recerca Consolidat 2009 SGR 1101).

## References

- Alvarez-Ponce D, Aguadé M, Rozas J. 2009. Network-level molecular evolutionary analysis of the insulin/TOR signal transduction pathway across 12 *Drosophila* genomes. *Genome Res.* 19(2):234–242.
- Andrés AM, de Hemptinne C, Bertranpetit J. 2007. Heterogeneous rate of protein evolution in serotonin genes. *Mol Biol Evol.* 24(12):2707–2715.
- Birney E, Clamp M, Durbin R. 2004. GeneWise and Genomewise. *Genome Res.* 14(5):988–995.
- Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol.* 17:540–552.
- Cork JM, Purugganan MD. 2004. The evolution of molecular genetic pathways and networks. *Bioessays* 26(5):479–484.
- Flowers JM, Sezgin E, Kumagai S, Duvernell DD, Matzkin LM, Schmidt PS, Eanes WF. 2007. Adaptive evolution of metabolic pathways in *Drosophila*. *Mol Biol Evol.* 24(6):1347–1354.
- Fraser HB, Hirsh AE, Steinmetz LM, Scharfe C, Feldman MW. 2002. Evolutionary rate in the protein interaction network. *Science* 296(5568):750–752.
- Gagneaux P, Varki A. 1999. Evolutionary considerations in relating oligosaccharide diversity to biological function. *Glycobiology* 9(8):747–755.
- Goslee S, Urban D. 2007. The ecodist package for dissimilarity-based analysis of ecological data. *J Stat Softw.* 22:7.
- Greenberg AJ, Stockwell SR, Clark AG. 2008. Evolutionary constraint and adaptation in the metabolic network of *Drosophila*. *Mol Biol Evol.* 25(12):2537–2546.
- Hahn MW, Kern AD. 2005. Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. *Mol Biol Evol.* 22(4):803–806.
- Hirsh AE, Fraser HB. 2001. Protein dispensability and rate of evolution. *Nature* 411(6841):1046–1049.
- Innan H, Kondrashov F. 2010. The evolution of gene duplications: classifying and distinguishing between models. *Nat Rev Genet.* 11(2):97–108.
- Jeong H, Mason SP, Barabási AL, Oltvai ZN. 2001. Lethality and centrality in protein networks. *Nature* 411(6833):41–42.

- Kim PJ, Lee DY, Jeong H. 2009. Centralized modularity of N-linked glycosylation pathways in mammalian cells. *PLoS One*. 4(10):e7317.
- Kim PM, Korbel JO, Gerstein MB. 2007. Positive selection at the protein network periphery: evaluation in terms of structural constraints and cellular context. *Proc Natl Acad Sci U S A*. 104(51):20274–20279.
- Lemos B, Bettencourt BR, Meiklejohn CD, Hartl DL. 2005. Evolution of proteins and gene expression levels are coupled in *Drosophila* and are independently associated with mRNA abundance, protein length, and number of protein-protein interactions. *Mol Biol Evol*. 22(5):1345–1354.
- Livingstone K, Anderson S. 2009. Patterns of variation in the evolution of carotenoid biosynthetic pathway enzymes of higher plants. *J Hered*. 100(6):754–761.
- Lu Y, Rausher MD. 2003. Evolutionary rate variation in anthocyanin pathway genes. *Mol Biol Evol*. 20(11):1844–1853.
- Notredame C, Higgins DG, Heringa J. 2000. T-Coffee: a novel method for multiple sequence alignments. *J Mol Biol*. 302:205–217.
- Ramsay H, Rieseberg LH, Ritland K. 2009. The correlation of evolutionary rate with pathway position in plant terpenoid biosynthesis. *Mol Biol Evol*. 26(5):1045–1053.
- Rausher MD, Miller RE, Tiffin P. 1999. Patterns of evolutionary rate variation among genes of the anthocyanin biosynthetic pathway. *Mol Biol Evol*. 16(2):266–274.
- Riley RM, Jin W, Gibson G. 2003. Contrasting selection pressures on components of the Ras-mediated signal transduction pathway in *Drosophila*. *Mol Ecol*. 12(5):1315–1323.
- Rocha EP. 2006. The quest for the universals of protein evolution. *Trends Genet*. 22:412–416.
- Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*. 13(11):2498–2504.
- Sokal RR, Rohlf FJ. 1995. *Biometry*, 3rd ed. New York: Freeman.
- Storey JD, Tibshirani R. 2003. Statistical significance for genomewide studies. *Proc Natl Acad Sci U S A*. 100:9440–9445.
- Swanson WJ, Nielsen R, Yang Q. 2003. Pervasive adaptive evolution in mammalian fertilization proteins. *Mol Biol Evol*. 20(1):18–20.
- Varki A. 2006. Nothing in glycobiology makes sense, except in the light of evolution. *Cell* 126(5):841–845.
- Varki A, Cummings RD, Esko JD, Freeze HH, Stanley P, Bertozzi CR, Hart GW, Etzler ME. 2009. *Essentials of Glycobiology*, 2nd ed. Cold Spring Harbor (NY): Cold Spring Harbor Laboratory Press.
- Vilella AJ, Severin J, Ureta-Vidal A, Heng L, Durbin R, Birney E. 2009. EnsemblCompara GeneTrees: complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res*. 19(2):327–335.
- Vitkup D, Kharchenko P, Wagner A. 2006. Influence of metabolic network structure and function on enzyme evolution. *Genome Biol*. 7(5):R39.
- Wilson AC, Carlson SS, White TJ. 1977. Biochemical evolution. *Annu Rev Biochem*. 46:573–639.
- Yamada T, Bork P. 2009. Evolution of biomolecular networks: lessons from metabolic and protein interactions. *Nat Rev Mol Cell Biol*. 10(11):791–803.
- Yang YH, Zhang FM, Ge S. 2009. Evolutionary rate patterns of the Gibberellin pathway genes. *BMC Evol Biol*. 9:206.
- Yang Z. 2006. *Computational molecular evolution*. Oxford: Oxford University Press.
- Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol*. 24(8):1586–1591.
- Yang Z, Bielawski JP. 2000. Statistical methods for detecting molecular adaptation. *Trends Ecol Evol*. 15(12):496–503.
- Yang Z, Nielsen R, Goldman N, Pedersen AM. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics* 155(1):431–449.