Positive selection at the protein network periphery: Evaluation in terms of structural constraints and cellular context

Philip M. Kim*†, Jan O. Korbel*†, and Mark B. Gerstein*†‡§

*Department of Molecular Biophysics and Biochemistry, ‡Department of Computer Science, and §Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT 06520

Communicated by Donald M. Engelman, Yale University, New Haven, CT, October 26, 2007 (received for review February 4, 2007)

Because of recent advances in genotyping and sequencing, human genetic variation and adaptive evolution in the primate lineage have become major research foci. Here, we examine the relationship between genetic signatures of adaptive evolution and network topology. We find a striking tendency of proteins that have been under positive selection (as compared with the chimpanzee) to be located at the periphery of the interaction network. Our results are based on the analysis of two types of genome evolution, both in terms of intra- and interspecies variation. First, we looked at single-nucleotide polymorphisms and their fixed variants, single-nucleotide differences in the human genome relative to the chimpanzee. Second, we examine fixed structural variants, specifically large segmental duplications and their polymorphic precursors known as copy number variants. We propose two complementary mechanisms that lead to the observed trends. First, we can rationalize them in terms of constraints imposed by protein structure: We find that positively selected sites are preferentially located on the exposed surface of proteins. Because central network proteins (hubs) are likely to have a larger fraction of their surface involved in interactions, they tend to be constrained and under negative selection. Conversely, we show that the interaction network roughly maps to cellular organization, with the periphery of the network corresponding to the cellular periphery (i.e., extracellular space or cell membrane). This suggests that the observed positive selection at the network periphery may be due to an increase of adaptive events on the cellular periphery responding to changing environments.

protein structure | network centrality | single-nucleotide change | copy number variant | structural variant

With the advent of genomic sequence data and, more recently, large-scale genetic variation data (1, 2), it has become possible to examine genes or genomic regions for signs of recent evolutionary adaptation in our genome, characterized as signatures of positive selection (3, 4). Typically, tests for positive selection predict adaptation by testing and rejecting the hypothesis of neutral mutation (5) or variation for a given genomic region.

Despite considerable advances in the field of genetics, the actual molecular relationship of recent evolutionary events with biophysical properties of associated proteins such as structural characteristics and network connectivity (i.e., protein interactions) has as yet not been studied in detail. Understanding the extent of recent mutations, polymorphisms, and adaptation beyond their effect on the gene level is crucial because most complex cellular processes only come about through the interplay and interactions of many different proteins. On the other hand, although recent proteomic surveys have suggested that proteins with many interaction partners are subject to considerable structural constraints, the connection with human genome variation has not yet been considered. Thus, by combining knowledge from evolution and biophysics, new conclusions on cause and effect of variation on molecular processes can be found.

Single base pair changes drift through the population after their emergence and are visible as single-nucleotide polymorphisms (SNPs) before becoming fixed as substitutions. A popular method to scan for positive selection is comparing the ratio of nonsynonymous to synonymous substitutions (known as the dN/dS ratio) with respect to another species, such as the chimpanzee (3). Fuelled by the emergence of large-scale sequence and SNP genotyping data, a number of studies have reported signs of recent adaptation for genes or larger regions in the human genome (6–11).

In addition the spectrum of variation in the human genome goes beyond SNPs: in particular, large-scale structural variants (i.e., kb up to Mb rearrangements) in the form of deletions, duplications, insertions, and inversions occur commonly in humans (12–15). Copy number variants (CNVs) (i.e., deletions and duplications) are the best-studied form of structural variation (12–15). They account for a major portion of intraspecies variation (15) and have been implicated in adaptive evolution (16). Similar to SNPs, CNVs are expected to drift through the population and upon fixation (by drift or selection) will be detectable in the genome as segmental duplications (SDs) (17).

Genomewide studies of evolutionary aspects of SNPs and CNVs have not yet been related to structural properties of the affected proteins and their position in the protein network; consequently, the relationship between recent adaptation events and the structure and evolutionary dynamics (i.e., rewiring of edges and addition of new nodes) of the interactome are unclear. Recently, initial versions of the human protein interaction network, or interactome, have been described following large-scale literature curation (18) and yeast-two-hybrid screens (19, 20). Here, we study signs of recent adaptation in terms of the human protein interaction network and protein structure. In particular, we provide evidence for proteins at the network periphery to be preferentially involved in recent or ongoing adaptive evolution, manifested in two complementary forms of molecular evolution; namely, single base pair mutations and segmental duplications. To investigate this trend, we do further analysis in terms of protein structure and population genetic variation. We find that we can rationalize it both in structural and cellular terms.

Results

Positive Selection on Single Base Pair Changes Occurs Preferentially at the Network Periphery. To assess whether evolutionary adaptation is biased to certain regions of the interactome, we initially focused

Author contributions: P.M.K. and J.O.K. contributed equally to this work; P.M.K., J.O.K., and M.B.G. designed research; P.M.K. performed research; P.M.K. and J.O.K. analyzed data; and P.M.K., J.O.K., and M.B.G. wrote the paper.

The authors declare no conflict of interest.

Freely available online through the PNAS open access option.

†To whom correspondence may be addressed. E-mail: pmkim@alum.mit.edu, jan.korbel@yale.edu, or mark.gerstein@yale.edu.

This article contains supporting information online at www.pnas.org/cgi/content/full/0710183104/DC1.

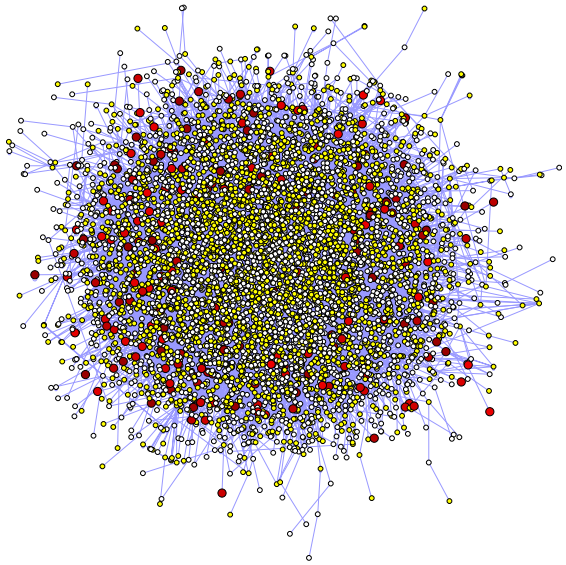

Fig. 1. The human protein interaction network and its connection to positive selection. Proteins likely to be under positive selection are colored in shades of red (light red, low likelihood of positive selection; dark red, high likelihood) (6). Proteins estimated not to be under positive selection are in yellow, and proteins for which the likelihood of positive selection was not estimated are in white (6).

on single base pair changes. We calculated two measures of topological centrality, betweenness centrality and degree centrality for all proteins in the human protein interaction network (21, 22). Briefly, the betweenness of a node is the number of shortest paths that pass through it and is, hence, a global measure of centrality. Conversely, the degree corresponds to its number of interaction partners and is a local measure. We then related these centrality statistics to signatures for positive selection based on a recent scan that used the dN/dS ratio test (6). For every protein, we compared its centrality with the likelihood ratio from the dN/dS test (roughly, the probability for positive selection) of the associated gene. Intriguingly, we observed the fraction of genes under recent positive selection to be considerably higher in the periphery of the network than in the center. Furthermore, the probability of a gene to be under positive selection significantly correlates with its centrality {both for betweenness and degree [see *Methods*, Figs. 1 and 2*A*, and supporting information (SI) Fig. 4], Spearman correlation $\rho = -0.06$, $P = 2.9e\text{-}05$ for betweenness, $\rho = -0.07$, $P = 6.7e\text{-}06$ for degree; this correlation is fairly weak but significant}. Put a different way, proteins that are likely to be under positive selection tend to be positioned at the network periphery, whereas proteins unlikely to have been positively selected recently are at the center: Proteins with dN/dS > 1 have an average betweenness centrality of 27,085 paths, whereas proteins with dN/dS ≤ 1 have an average betweenness centrality of about twice that much. This difference is highly significant with a P value of $2.3e\text{-}05$ (Fig. 2*C* and SI Fig. 4 for degree).

To ensure that this observation is not a result of inherent data biases, we examined whether it would hold up to our varying a number of factors. Because current interaction networks are incomplete and may suffer from biases, we examined a number of different networks. We find the trend to be present in many interaction datasets that are based on both literature curation efforts and high-throughput screens (SI Table 6) (18, 20). Because these datasets have small overlap among each other (23), it is reasonable to assume that in a complete interaction network, one would observe the same result. Furthermore, the trend is present in two different estimations of positive selection based on the dN/dS ratio test (6, 10). Yet another influencing factor that might affect

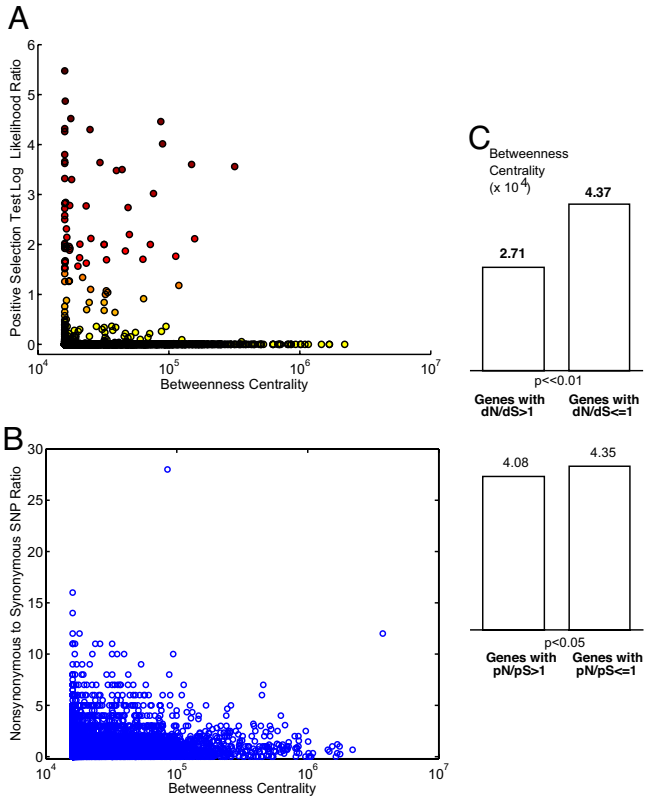

Fig. 2. Relationship of protein network centrality and single-nucleotide changes. (*A*) The periphery of the human interactome is strongly enriched for genes under positive selection. Shown is the correlation of the likelihood to be positively selected (6) and betweenness centrality (18). Dots are colored according to the same scheme as in Fig. 1. As expected for a highly significant Spearman rank correlation, almost all dots are near the x axis for high betweenness centralities, whereas high probabilities for positive selection are only observed at low betweenness centralities (Spearman $\rho = -0.06$, significant at $P = 1.2e\text{-}06$). (*B*) The periphery of the human interaction network is more variable on the protein sequence level. Shown is the ratio of nonsynonymous to synonymous SNPs vs. network centrality. A higher ratio (which corresponds to variability at the protein sequence level) tends to occur at the network periphery (Spearman $\rho = -0.1$, significant at $P = 4.0e\text{-}04$). (*C Upper*) Betweenness centrality of genes with some likelihood of being under positive selection (with a log-likelihood ratio >0) vs. all other genes. (*C Lower*) Betweenness centrality of genes with a high ratio of nonsynonymous to synonymous SNPs vs. genes with a low ratio of nonsynonymous to synonymous SNPs. The significance level of the differences is given as the Wilcoxon rank sum P value between the bars.

our result is the known anticorrelation of mutational rate and gene expression (24). Previously reported in yeast, we found an equivalent relationship also in humans (see SI Fig. 5) and furthermore observed a similar (possibly related) correlation for genes under adaptation: i.e., most positively selected genes tend to be expressed at a low level (Table 1). Conversely, central proteins tend to be expressed at a higher level than peripheral ones (see SI Fig. 6). We thus calculated partial correlations to rule out the possibility of gene expression biases that may have influenced the observed trends (see *Methods*). Indeed, both gene expression and network topology show independent and highly significant relationships with the likelihood of positive selection (Table 1). This shows that positively selected genes, aside from being expressed at low levels, are strongly enriched at the protein network periphery. All of these findings suggest that the trend is unlikely to stem from inherent biases in the data but is likely to be due to the constraints imposed by interactions on protein structures or the cellular context.

Table 1. Spearman rank correlation and partial correlation of gene expression, betweenness centrality, and positive selection likelihood

Parameter	ρ	P
Correlation with positive selection		
Network centrality, ρ_{bp}	−0.06	$\ll$0.01
Gene expression, ρ_{gp}	−0.04	0.01
Partial correlation with positive selection		
Network centrality, $\rho_{bp\|g}$	−0.06	$\ll$0.01
Gene expression, $\rho_{gp\|b}$	−0.03	0.06

ρ_{gp}, rank correlation coefficient of gene expression and positive selection likelihood; $\rho_{gp|b}$, partial rank correlation coefficient of gene expression and positive selection while controlling for betweenness; ρ_{gp}, rank correlation coefficient of betweenness centrality and positive selection likelihood, $\rho_{gp|b}$, partial rank correlation coefficient of betweenness centrality and positive selection while controlling for gene expression.

Features of Positively Selected Sites in 3D Protein Structures. A straightforward structural explanation for the preference of positive selection for the network periphery is stronger 3D-structural constraint on central proteins in the interaction network. This constraint (resulting from more interaction partners) would cause the proteins to evolve more slowly and be less likely to show signs of positive selection (e.g., when assessed by the dN/dS ratio test). Noncentral nodes, on the other hand, should be under relaxed constraint, and the enrichment of adaptation at the network periphery may be due to the associated increased variability. To investigate this influence of relaxation of structural constraints, we sought to analyze the structural features of the sites in question. Structural constraint would have a significant effect if positively selected amino acids are preferentially positioned at the protein surface, which should underlie different constraints for peripheral proteins as opposed to central ones (hubs) (25, 26)—in particular for hubs with multiple interfaces involved in protein complexes. Indeed, we found that residues positioned on a protein's accessible surface are under significantly less evolutionary constraint [having a substantially higher dN/dS ratio (Tables 2–4)]. Likewise, the average relative surface accessibility of sites that have nonsynonymous nucleotide differences when compared with chimpanzee genes is significantly higher than sites that only have synonymous (silent) differences [suggesting that nonsynonymous sites are enriched on the protein surface (Tables 2–4)]. These results are consistent with earlier studies (e.g., refs. 27 and 28). However, when we examined the nonsynonymous sites more closely, we observed another trend. We split all sites with nonsynonymous substitutions into two groups: Those that are likely to be under positive selection and those that are not—i.e., one group with all nonsynonymous sites in proteins that show dN/dS $>$ 1 in human–chimpanzee alignments and a second group with all nonsynonymous sites in proteins that have dN/dS $\leq$ 1. We found that the average relative surface accessibility is significantly lower for the former group (Tables 2–4). This result indicates that mutations that lead to a fitness advantage are likely to be somewhat buried and may lie in clefts. Hence, they would have a higher impact on the protein structure and function than neutral mutations. This is reasonable because mutations at completely exposed sites would likely not have a larger impact on the proteins' function. In line with these findings, amino acid changes in central proteins are significantly more exposed than those changes in peripheral proteins, indicating that strong functional and structural constraints would favor mutations that are exposed and have a lighter effect on overall protein structure and function (Fig. 3*C*).

Table 2. Surface-exposed sites are under significantly less evolutionary constraint than buried sites

Sites	Exposed	Buried
dN/dS	*0.49*	0.35

The dN/dS ratio is shown for sites with an average relative surface exposure [calculated by SABLE (42)] of >70% (exposed sites) and <50% (buried sites). Italics indicates significant differences.

Table 3. Sites with nonsynonymous mutations are more exposed than sites with synonymous (silent) mutations

Sites	Synonymous	Nonsynonymous
Average ASA	2.26	*2.66*

The average surface index (from SABLE) of all sites with nonsynonymous nucleotide differences (between human and chimpanzee gene sequences) are shown vs. the average surface index of sites with synonymous nucleotide differences. Italics indicates significant differences. ASA, accessible surface area (SABLE estimate).

Correspondence of the Cellular Periphery with the Interactome Periphery. Another explanation for our results of positive selection at the network periphery is that adaptive evolution may preferentially occur there—i.e., there may be an ongoing need for adaptation. That is, in contrast to the more ancient network center, which is responsible for conserved essential functions, the periphery of the network may still be more adaptable to changing environments. In this sense, the network periphery would functionally correspond to the cellular periphery. This correspondence would represent a separate and complementary explanation for the trends observed here. Positively selected genes (6–8) have been shown to be significantly enriched in environment response genes. It is hence reasonable to hypothesize that they would be located at the "periphery" to interact with the changing environment. We have shown thus far that they are indeed located at the periphery in a network-topology sense. However, a more straightforward notion is the periphery in a cellular context. In this sense, extracellular proteins can be considered as the "natural periphery" of the proteome, and indeed, they have both a lower average degree and betweenness than proteins belonging to other cellular components (Table 5). Moreover, the average centrality statistics of proteins belonging to various cellular components appears to follow our intuition of central and peripheral subcellular locales. Furthermore, when examining cellular component gene ontology (GO) terms (which describe the subcellular localization of proteins) for enrichment in positively selected proteins, only the GO terms of "peripheral" cellular components (e.g., "extracellular space" and "extracellular region") are significantly enriched [with a false discovery rate of <0.06 (see *Methods*)]. Therefore, part of our observed trend may be explained by the fact that the network periphery corresponds to the cellular periphery and is responsible for mediating interactions with the environment. Although some GO categories

Table 4. Comparison of nonsynonymous sites in proteins under positive selection and peripheral and central proteins

Nonsynonymous sites	dN/dS $>$ 1	dN/dS $\leq$ 1	Peripheral	Central
Average ASA	2.26	*2.66*	2.45	*2.79*

Comparing different nonsynonymous sites, those in proteins that are estimated to be under positive selection are less exposed than the ones for other proteins. Likewise, those in peripheral proteins are less exposed than those in central proteins. The average surface index of all nonsynonymous sites in proteins with dN/dS $>$ 1 is shown vs. all nonsynonymous sites in proteins with dN/dS $\leq$ 1. Likewise, The average surface index of all nonsynonymous sites in proteins with a betweenness of <10,000 vs. all nonsynonymous sites in proteins with a betweenness of >10,000. Italics indicates significant differences. ASA, accessible surface area (SABLE estimate).

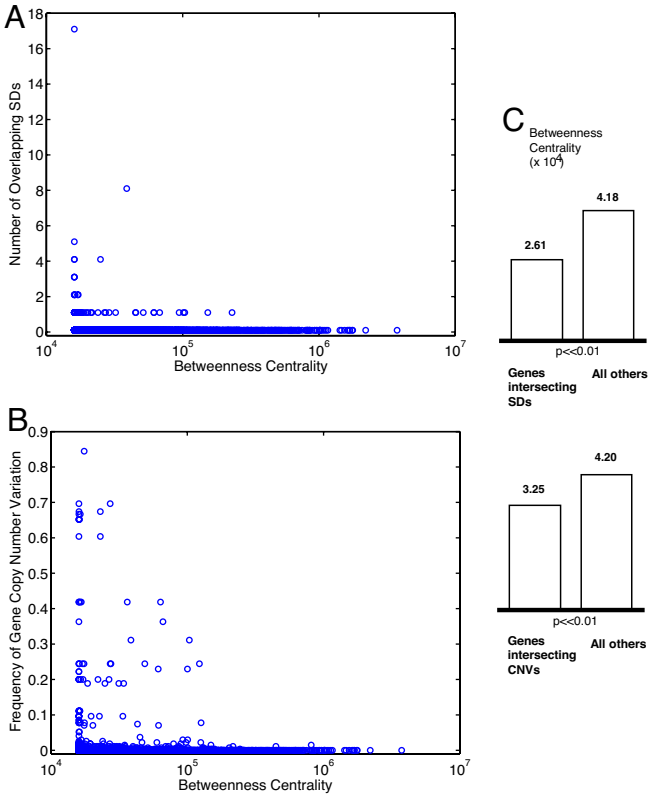

Fig. 3. Relationship of protein network centrality and changes in genetic copy number. (*A*) Correlation of the number of overlapping SDs of each gene with the betweenness centrality of the associated protein (Spearman $\rho = -0.04$, significant at $P = 3.3e\text{-}03$). (*B*) The periphery of the human interaction network is more variable on the level of genome rearrangements. Shown is the frequency of CNVs that intersect a given gene vs. the corresponding protein's network centrality (Spearman $\rho = -0.03$, significant at $P = 0.002$). (*C Upper*) Betweenness centrality of genes that intersect with at least one SD vs. centrality of all other genes. (*C Lower*) Betweenness centrality of genes that intersect with at least one CNV vs. the centrality of all other genes. The significance level of the differences is given as the Wilcoxon rank sum *P* value between the bars.

preferentially occur at the network periphery, for a sizeable number of tested categories the significant correlations between positive selection and network centrality/betweenness remain even when only proteins within the category are analyzed (SI Table 7).

Proteins on the Network Periphery Have a Higher Propensity for Nonsynonymous SNPs. In summary, we have shown that the preference of positive selection for the network periphery may be accounted for by two complementary explanations: structural constraint and cellular context. Next, we examine the relationship of population genetic variability and protein networks. Relaxed constraint would manifest itself in an increase of genetic variability at the network periphery, comparable in magnitude to the preference of positively selected genes at the periphery. One measure of genetic variability at the protein coding level is given by the ratio of nonsynonymous (having an effect on the protein sequence) to synonymous (silent with respect to the sequence) SNPs, known as the pN/pS ratio. This ratio is analogous to the dN/dS ratio, but because it measures intraspecies variation, it can be viewed as a measure of variability. We found that there generally is a higher ratio of nonsynonymous to synonymous SNPs at the network periphery [Spearman correlation $\rho = -0.1$, $P = 4.0e\text{-}04$ (Fig. 2*B*)]. This indeed suggested stronger evolutionary constraint for proteins at the network center, resulting in stronger negative selection and in turn removing a larger proportion of nonsynonymous SNPs. However, we note that the trend for the pN/pS ratio is weaker than for the dN/dS ratio (Fig. 2*C*). This suggests that if only relaxation of structural constraints were taken into account, the observed trends may be only insufficiently explained.

Segmental Duplications Preferentially Occur on the Network Periphery. Evolution of protein coding genes by single base pair mutations is only one of many evolutionary processes. Hence, we examined whether other mechanisms would also exhibit a preference for proteins on the network periphery. In particular, we initially focused on SDs, duplications that presumably have been fixed (a subset of SDs in the human reference genome may correspond to unrecognized CNVs). Namely, we found that SDs have a preference to be associated with genes positioned at the network periphery [for betweenness, Spearman correlation $\rho = -0.04$, $P = 4.6e\text{-}03$ (Fig. 3*A*); for degree, Spearman correlation $\rho = -0.04$, $P = 3.3e\text{-}03$ (SI Fig. 7)]. In fact, the more SDs intersected with a given gene in the human reference genome, the stronger was the preference for the encoded protein to be positioned in the periphery of the protein network. Genes intersecting with SDs have an average betweenness centrality of 26,119, whereas genes that do not intersect with SDs have an average betweenness centrality of 41,775 [rank sum significance of $P = 4.8e\text{-}04$ (Fig. 3*C*); for degree, see SI Fig. 7]. This agrees with previous findings in yeast; i.e., that duplication events are more frequent for proteins with low network connectivity (29), which at least in part may be caused by the dosage-sensitivity of components of large protein complexes (30).

Analysis of Copy Number Variants Provides Additional Evidence for Adaptive Events at the Network Periphery. Analogous to our comparison of SNPs and fixed differences above, we investigated a measure of intraspecies variation and its relationship to network centrality in comparison to the results found for SDs. SDs are the result of fixation of CNVs, in particular those corresponding to duplications (here referred to as "Gain-CNVs"). Given this, we analyzed the relationship of CNVs to the protein network. Our analysis is based on the assumption that relating both SDs and CNVs to the network topology may enable us to recognize (or reject) signs of recent adaptation. In particular, the prevalence of CNVs in a given genomic region can be viewed as a measure of its variability in terms of chromosomal rearrangements: A region having a high incidence of CNVs is likely to be more variable than a region having a low incidence. Variability can potentially be influenced by a number of factors, such as genomic stability, different propensities for occurrences of double-strand breaks, or recombination events (31). To examine whether the prevalence of

Table 5. Gene Ontology (GO Slim) cellular component terms and association of network periphery to positive selection

GO Slim cellular component	Extracellular region	Membrane	Cytoplasm	Nucleus	Chromosome	All
Average degree	5.89	6.51	8.07	8.62	10.22	6.85
Average betweenness	37,857	40,333	51,537	50,026	55,178	41,617

Shown is the average degree and betweenness of proteins that are annotated to the GO cellular component terms. Also shown are the Spearman correlation of the betweenness centrality with the likelihood ratio of positive selection when only considering genes from this particular GO term. Peripheral cellular components also tend to lie on the network periphery.

SDs to occur in the network periphery is merely a result of increased variability, we examined whether CNVs as well would operate mostly on peripheral proteins. If increased variability (due to relaxed constraints) were the only reason, we would expect the same degree of enrichment of CNVs at the network periphery. After mapping all genes overlapping CNVs to the protein interaction network, we find that CNVs (Gain- as well as Loss-CNVs, or deletions) have a significant but much less pronounced tendency than SDs for operating preferentially on peripheral proteins [for betweenness (see Fig. 3 *B* and *C*), $\rho = -0.03$, significant at $P = 0.003$; for degree (see SI Fig. 7), $\rho = -0.03$, significant at $P = 0.002$]. This suggests that the preference of SDs to operate on peripheral genes is not simply a result of increased variability or relaxed constraint at the network periphery. Taken together, we find additional support for ongoing preferential fixation of copy number variants at the network periphery related to evolutionary adaptation. Also note that the genes intersecting segmental duplications have been shown to be significantly enriched in environmental interactions (16, 32, 33). It would hence be reasonable that they would be located at the (cellular and network) periphery.

Discussion

We have presented evidence for a preference of recent and ongoing adaptive events for the periphery of the human protein interaction network. We present two possible explanations for this trend. First, a structural analysis shows a preference of positive selected sites for presumed functional clefts on the protein surface, which indicates that structural constraints would lead to a depletion of these at central proteins; conversely, at peripheral proteins, these constraints would be relaxed. Second, we find a correspondence of the cellular periphery with the network periphery and a preference of positively selected genes to belong to both the cellular and the network periphery. Together with an enrichment of functions that relate to environmental interactions, this result indicates that a stronger exposure to the environment would cause a stronger need for adaptation at peripheral proteins.

We examine adaptive evolution in two guises: Protein evolution by single base pair changes and genome evolution through segmental duplications. For both of these mechanisms, we have looked at fixed differences and intraspecies variation. The effect of relaxation of constraint would be visible in both fixed differences and intraspecies variation, whereas adaptive evolutionary events through environmental exposure are less likely to have an effect on intraspecies variation. In both the single base pair and the large-scale duplication cases, we observe that the preference for the network periphery is stronger for the fixed differences than the intraspecies variability. We believe that this result indicates that the relaxation of structural constraints and environmental pressure are complementary explanations for the propensity of peripheral proteins to be under positive selection or part of segmental duplications.

Among many interesting examples of proteins at the network periphery that may be under positive selection are the protein encoded by the CHRNA5 (ENSG00000169684) gene, neuronal acetyl-choline receptor subunit α-5; this integral membrane protein is involved in neuronal processes likely to be under ongoing adaptation, and CHRNA5 was recently associated with several cognitive performance criteria (34). Another protein, the Ficolin-3 protein encoded by the FCN3 gene (ENSG00000142748), is a secreted protein that exerts lectin activity and is presumably involved in innate immunity through binding to bacterial lipopolysaccharides (35).

Recent studies have examined disease proteins and essential proteins in the context of the interaction network (18, 36). Not surprisingly, essential proteins tend to lie in the center of the network, which is consistent with our results—the core of the network is conserved, essential, and in no further need of adaptation. Interestingly, Goh *et al.* (36) provide evidence that proteins involved in genetic diseases show little preference for either the center or the periphery. This is also consistent with our results. The diseases in their dataset that are of a genetic nature (e.g., leukemia, etc.)—i.e., they are "intrinsic" diseases and are hence not involved with the environment—are also not that likely to be involved in adaptive evolution. Conversely, proteins that are involved in dealing with externally caused diseases (e.g., proteins involved in immune response) are likely to be on the cellular periphery.

Similar trends that relate topology to variation may be expected in other types of biological networks—for instance, in regulatory networks that involve microRNAs, although the lack of codons in their genes would require different types of adaptive selection tests. Moreover, the general notion of adaptation and variation on the periphery and constraint at the center obviously has analogies in other types of networks—e.g., innovation coming in from the borders of social networks. The parallels are particularly clear with respect to security considerations in computer networks. Computers (nodes in the network, analogous to proteins) tend to be connected in local networks (analogous to cells) that are in turn interconnected into larger networks (the environment)—e.g., the internet. Computers at the periphery of an internal network are patched much more frequently to protect them against security threats, similar to the process of genetic mutation favored by positive selection (37). Conversely, computers that sit at the very center of an internal network are often large servers under heavy use, which puts great constraints on the ease with which they can be updated. This situation is analogous to what we observe in the protein interaction network.

Methods

Relationship of Network Structure and Positive Selection. Interaction data were combined from the Human Protein Reference Database (HPRD) (18), which is based on small-scale studies that were curated from the literature, and from two recent high-throughput yeast-two-hybrid screens (19, 20). The combined network contained a total of 30,239 interactions among 8,383 proteins. As a measure of how central each protein is in the network, both the betweenness [the number of shortest paths running through a node (21)] and the degree [the number of interaction partners (22)] were calculated. Positive selection data were gathered from two recent scans using the dN/dS ratio test (6, 10). The screens calculated the likelihood ratio of positive selection for 8,079 and 7,645 genes, respectively. Significant positive deviations from neutrality (dN/dS = 1) are a conservative measure of positive selection. Briefly, the reasoning for this notion is that during a period of neutral evolution, the rate of synonymous or nonsynonymous mutations should be equal. If there are more nonsynonymous than synonymous mutations, at least some of the nonsynonymous mutations were fixed preferentially, which indicates positive selection. (For a more detailed description, see refs. 3 and 38.) Nielsen *et al.* (6) used a likelihood ratio test to infer likelihood ratios from the dN/dS data. This method detects positive selection at loci that have been under repeated mutational selection pressure. Interaction data and positive likelihood data were mapped to Ensembl gene IDs (39). A total of 3,727 genes were present in both interaction data and positive selection scan (6); on these, Spearman rank correlations were calculated. To exclude the possibility of a gene expression prebias, we also gathered expression data from the human expression atlas (40). We used the average of all robust multiarray average (RMA) (41) expression values from Affymetrix microarray experiments across multiple tissues and also calculated expression breadth across tissues by counting the number of tissues in which a certain gene is in above the 80th percentile in RMA values. We then inferred the relationship between positive selection likelihood and betweenness by computing partial correlation coefficients. Partial correlation corresponds to the correlation between two variables while controlling for a third variable. We computed the partial correlation of the ranks. The partial correlation among betweenness and positive selection while controlling for expression was still significant, demonstrating that network centrality has an effect on positive selection independent from gene expression.

Relationship of Network Structure and SNPs. dbSNP was used as the source for SNP data. SNP locations, annotations into nonsynonymous and synonymous, and its mapping to Ensembl gene IDs was downloaded from Ensembl.

Calculation of Protein Surface Index of Mutated Sites. The predicted surface accessibility of each residue was calculated by using the relative surface accessibility predictor SABLE (42). The mutated sites were identified by using the translations of the nucleotide alignments of human and chimpanzee genes by

Nielsen (6). For each protein, all surface indices were averaged and compared with the likelihood ratio of positive selection.

Relationship of Network Structure and SDs. SDs were downloaded from the Segmental Duplication Database [http://humanparalogy.gs.washington.edu (43)], a database reporting recent duplications according to the criterion >90% sequence identity and >1 kb length. For each SD, all Ensembl (39) genes annotated as being affected by a SD (including partial as well as full overlaps with given coding regions) were presumed to be associated with it. A total of 25,318 SDs were analyzed, intersecting 2,173 genes. For genes that were annotated as affected by more than one SD, we counted the number of SDs intersecting each gene and refer to it as the number of SDs affecting the gene.

Relationship of Network Structure and Variation. The locations of CNVs were downloaded from the Database of Genomic Variants [http://projects.tcag.ca/variation (13)]. We focused on the set of Redon *et al.* (15) generated by using genomewide high-resolution SNP genotyping arrays, which represents the highest-resolution comprehensive CNV mapping carried so far. CNVs are classified as ''Gain-CNVs'' and ''Loss CNVs'' based on observed array signals (15). While it is thought that Gain-CNVs correspond to amplifications (i.e., an increase in copy number) and Loss-CNVs to deletions (copy number decrease), it is also known that because of a number of confounding factors [such as the control individual(s) used in a DNA microarray experiments], this correlation is not perfect. For each CNV, all Ensembl genes that are annotated as being affected by a CNV (including partial and full overlaps of given coding regions) were presumed to be associated with the CNV. A total of 406 Gain-CNVs and 697 Loss-CNVs were analyzed, intersecting with 1,649 and 1,443 Ensembl genes, respectively. Frequencies were estimated by dividing the number of times a CNV was observed in a set of experiments by the total number of studied samples.

Analysis of GO Terms. All analyzed genes were mapped to terms of the GOA GO Slim ontology (44), obtained from www.ebi.ac.uk. For the proteins assigned to each GO term, rank correlations between dN/dS likelihood ratios and network parameters were calculated separately. For the enrichment of GO terms in peripheral and central proteins under positive selection, GoMiner was used (45). We report only GO terms that have significant enrichment after applying a multiple hypothesis testing correction, and that have a false discovery rate of <0.06.

Network Visualization. The human interactome in Fig. 1*A* was drawn with the visualization package Cytoscape (http://cytoscape.org). The layout was done automatically by using a spring-embedding algorithm. Thereby, node order (whether coinciding nodes are visible in the front or invisible/covered in the back) was random (Cytoscape default). After mapping the positive selection likelihoods (6) to the nodes, the trend of positive selection at the periphery was clearly visible, despite the fact that the layout algorithm did not optimize according to betweenness. However, high betweenness nodes tend to get put in the center of the graph because it usually connects a number of larger clusters. Putting them on the outside would also lead to a large increase in potential energy.

Complete data files from our analysis are available at www.gersteinlab.org/proj/netpossel.

ACKNOWLEDGMENTS. We thank K. Kidd, A. Urban, S. Weissman, and M. Snyder for valuable suggestions. We also thank the dataset producers. This work was supported by the National Institutes of Health. J.O.K. was supported by the European Union Sixth Framework Programme.

1. International HapMap Consortium (2005) *Nature* 437:1299–1320.
2. Chimpanzee Sequencing and Analysis Consortium (2005) *Nature* 437:69–87.
3. Nielsen R (2005) *Annu Rev Genet* 39:197–218.
4. Bamshad M, Wooding SP (2003) *Nat Rev Genet* 4:99–111.
5. Kimura M (1979) *Sci Am* 241:98–100, 102, 108 passim.
6. Nielsen R, Bustamante C, Clark AG, Glanowski S, Sackton TB, Hubisz MJ, Fledel-Alon A, Tanenbaum DM, Civello D, White TJ, *et al.* (2005) *PLoS Biol* 3:e170.
7. Voight BF, Kudaravalli S, Wen X, Pritchard JK (2006) *PLoS Biol* 4:e72.
8. Bustamante CD, Fledel-Alon A, Williamson S, Nielsen R, Hubisz MT, Glanowski S, Tanenbaum DM, White TJ, Sninsky JJ, Hernandez RD, *et al.* (2005) *Nature* 437:1153–1157.
9. Carlson CS, Thomas DJ, Eberle MA, Swanson JE, Livingston RJ, Rieder MJ, Nickerson DA (2005) *Genome Res* 15:1553–1565.
10. Clark AG, Glanowski S, Nielsen R, Thomas PD, Kejariwal A, Todd MA, Tanenbaum DM, Civello D, Lu F, Murphy B, *et al.* (2003) *Science* 302:1960–1963.
11. Akey JM, Eberle MA, Rieder MJ, Carlson CS, Shriver MD, Nickerson DA, Kruglyak L (2004) *PLoS Biol* 2:e286.
12. Sebat J, Lakshmi B, Troge J, Alexander J, Young J, Lundin P, Maner S, Massa H, Walker M, Chi M, *et al.* (2004) *Science* 305:525–528.
13. Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, Qi Y, Scherer SW, Lee C (2004) *Nat Genet* 36:949–951.
14. Tuzun E, Sharp AJ, Bailey JA, Kaul R, Morrison VA, Pertz LM, Haugen E, Hayden H, Albertson D, Pinkel D, *et al.* (2005) *Nat Genet* 37:727–732.
15. Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, Fiegler H, Shapero MH, Carson AR, Chen W, *et al.* (2006) *Nature* 444:444–454.
16. Nguyen DQ, Webber C, Ponting CP (2006) *PLoS Genet* 2:e20.
17. Bailey JA, Eichler EE (2006) *Nat Rev Genet* 7:552–564.
18. Gandhi TK, Zhong J, Mathivanan S, Karthick L, Chandrika KN, Mohan SS, Sharma S, Pinkert S, Nagaraju S, Periaswamy B, *et al.* (2006) *Nat Genet* 38:285–293.
19. Stelzl U, Worm U, Lalowski M, Haenig C, Brembeck FH, Goehler H, Stroedicke M, Zenkner M, Schoenherr A, Koeppen S, *et al.* (2005) *Cell* 122:957–968.
20. Rual JF, Venkatesan K, Hao T, Hirozane-Kishikawa T, Dricot A, Li N, Berriz GF, Gibbons FD, Dreze M, Ayivi-Guedehoussou N, *et al.* (2005) *Nature* 437:1173–1178.
21. Freeman LC (1977) *Sociometry* 40:35–41.
22. Albert R, Jeong H, Barabasi AL (2000) *Nature* 406:378–382.
23. Mathivanan S, Periaswamy B, Gandhi TK, Kandasamy K, Suresh S, Mohmood R, Ramachandra YL, Pandey A (2006) *BMC Bioinformatics* 7(Suppl 5):S19.
24. Pal C, Papp B, Lercher MJ (2006) *Nat Rev Genet* 7:337–348.
25. Teichmann SA (2002) *J Mol Biol* 324:399–407.
26. Kim PM, Lu L, Xia Y, Gerstein M (2006) *Science* 314:1938–1941.
27. Valdar WS, Thornton JM (2001) *Proteins* 42:108–124.
28. Walker DR, Bond JP, Tarone RE, Harris CC, Makalowski W, Boguski MS, Greenblatt MS (1999) *Oncogene* 18:211–218.
29. Prachumwat A, Li WH (2006) *Mol Biol Evol* 23:30–39.
30. Papp B, Pal C, Hurst LD (2003) *Nature* 424:194–197.
31. Zhou Y, Mishra B (2005) *Proc Natl Acad Sci USA* 102:4051–4056.
32. Infante JJ, Dombek KM, Rebordinos L, Cantoral JM, Young ET (2003) *Genetics* 165:1745–1759.
33. Dunham MJ, Badrane H, Ferea T, Adams J, Brown PO, Rosenzweig F, Botstein D (2002) *Proc Natl Acad Sci USA* 99:16144–16149.
34. Rigbi A, Kanyas K, Yakir A, Greenbaum L, Pollak Y, Ben-Asher E, Lancet D, Kertzman S, Lerer B (2007) *Genes Brain Behav*, 10.1111/j.1601–183X.2007.00329.
35. Tsujimura M, Miyazaki T, Kojima E, Sagara Y, Shiraki H, Okochi K, Maeda Y (2002) *Clin Chim Acta* 325:139–146.
36. Goh KI, Cusick ME, Valle D, Childs B, Vidal M, Barabasi AL (2007) *Proc Natl Acad Sci USA* 104:8685–8690.
37. Cheswick WR (1990) *Proceedings of the USENIX Summer 1990 Conference* (USENIX, Berkeley, CA), pp 233–237.
38. Kreitman M (2000) *Annu Rev Genomics Hum Genet* 1:539–559.
39. Birney E, Andrews D, Caccamo M, Chen Y, Clarke L, Coates G, Cox T, Cunningham F, Curwen V, Cutts T, *et al.* (2006) *Nucleic Acids Res* 34:D556–D561.
40. Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, Zhang J, Soden R, Hayakawa M, Kreiman G, *et al.* (2004) *Proc Natl Acad Sci USA* 101:6062–6067.
41. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP (2003) *Biostatistics* 4:249–264.
42. Adamczak R, Porollo A, Meller J (2004) *Proteins* 56:753–767.
43. Bailey JA, Gu Z, Clark RA, Reinert K, Samonte RV, Schwartz S, Adams MD, Myers EW, Li PW, Eichler EE (2002) *Science* 297:1003–1007.
44. Camon E, Magrane M, Barrell D, Lee V, Dimmer E, Maslen J, Binns D, Harte N, Lopez R, Apweiler R (2004) *Nucleic Acids Res* 32:D262–D266.
45. Zeeberg BR, Qin H, Narasimhan S, Sunshine M, Cao H, Kane DW, Reimers M, Stephens RM, Bryant D, Burt SK, *et al.* (2005) *BMC Bioinformatics* 6:168.