# Protein Structural Modularity and Robustness Are Associated with Evolvability

Mary M. Rorick*,[1,2], and Günter P. Wagner[2,3]

[1]Department of Genetics, Yale University, New Haven, Connecticut

[2]Yale Systems Biology Institute, Orange, Connecticut

[3]Department of Ecology and Evolutionary Biology, Yale University, New Haven, Connecticut

*Corresponding author: E-mail: mmrorick@gmail.com.

**Accepted:** 8 May 2011

## Abstract

Theory suggests that biological modularity and robustness allow for maintenance of fitness under mutational change, and when this change is adaptive, for evolvability. Empirical demonstrations that these traits promote evolvability in nature remain scant however. This is in part because modularity, robustness, and evolvability are difficult to define and measure in real biological systems. Here, we address whether structural modularity and/or robustness confer evolvability at the level of proteins by looking for associations between indices of protein structural modularity, structural robustness, and evolvability. We propose a novel index for protein structural modularity: the number of regular secondary structure elements (helices and strands) divided by the number of residues in the structure. We index protein evolvability as the proportion of sites with evidence of being under positive selection multiplied by the average rate of adaptive evolution at these sites, and we measure this as an average over a phylogeny of 25 mammalian species. We use contact density as an index of protein designability, and thus, structural robustness. We find that protein evolvability is positively associated with structural modularity as well as structural robustness and that the effect of structural modularity on evolvability is independent of the structural robustness index. We interpret these associations to be the result of reduced constraints on amino acid substitutions in highly modular and robust protein structures, which results in faster adaptation through natural selection.

**Key words:** modularity, designability, contact density, evolvability, protein evolution, robustness.

## Introduction

The extensive robustness of biological systems has long fascinated biologists. Robustness can be defined as the tendency for a system to maintain functionality under perturbation. Here, we will specifically concern ourselves with robustness under mutational perturbation because it is heritable change that is most immediately relevant to evolvability, which is the ability to respond to positive selection (Wagner and Altenberg 1996; Pigliucci 2008; Wagner 2008). Although robustness can, in theory, stifle adaptation under certain circumstances (i.e., under "neutral confinement" where mutational change does not cause significant phenotypic change) (Ancel and Fontana 2000; Sumedha et al. 2007; Cowperthwaite et al. 2008; Draghi et al. 2010), it generally confers evolvability to living systems because it allows them to undergo innovative modification without losing functionality (i.e., because adaptation is not typically limited by the extent to it can change via mutation, but rather, by the extent to which mutational change creates useful, nonlethal phenotypic variation) (Wagner 2005; Wagner 2008). Robustness also serves to maintain high fitness under conditions of random genetic and environmental noise (Wagner et al. 1997; Gibson and Wagner 2000; Meiklejohn and Hartl 2002; Wagner 2005). Modularity—which we define as the clustering of epistatic interactions—is an important form of robustness because it limits the number of system components that are affected by a given perturbation (Wagner and Altenberg 1996; Wagner 1996; Ancel and Fontana 2000; Fontana 2002; Kitano 2004; Bhattacharyya et al. 2006; Wagner et al. 2007). Diverse hypotheses for the origin of modularity have been proposed, and there has yet to be agreement on a final answer to this question (Lipson et al. 2002; Gardner and Zuidema 2003; Force et al. 2005; Misevic et al. 2006; Lynch 2007; Wagner et al. 2007). Similar to the case of robustness, while modularity

can be linked to reduced evolvability in some specific scenarios (Ancel and Fontana 2000; Hansen 2002; Griswold 2006), the consensus is that it generally facilitates adaptive change (Wagner and Altenberg 1996; Gerhart and Kirschner 1997; Bogarad and Deem 1999; Hartwell et al. 1999; Yang 2001; Cui et al. 2002; Xia and Levitt 2002; Beldade and Brakefield 2003; Bhattacharyya et al. 2006; Chen and Dokholyan 2006; Franz-Odendaal and Hall 2006; Pereira-Leal et al. 2006). Being a form of robustness, we can make the prediction that modularity should confer evolvability to a system. The goal of this study is to test this prediction, as well as the predicted connection between robustness and evolvability, at the level of proteins. To do this, we look for whether indices of protein structural modularity and protein structural robustness correlate with a protein evolvability index.

Indexing modularity in biological systems is not a simple task, despite the fact that biological systems—and proteins in particular—are nonrandomly modular (Schlosser and Wagner 2004; del Sol et al. 2009) and that modularity seems to increase through evolutionary time (Bonner 1988). In this study, we measure protein structural modularity by assessing the density of helices and β-sheet strands. Protein structural robustness can be indexed via contact density (England and Shakhnovich 2003), which is the average number of contacts a residue makes with other residues in the protein structure. It correlates strongly with protein designability, which is the number of protein sequences that stably fold into a given structure. Designability is an important predictor of the number of mutations a structure can tolerate, and so it is a good indicator of protein mutational robustness for relatively structured proteins (Li et al. 1996). It also seems to be important for the maintenance of stability and folding over the course of long periods of protein evolution (Govindarajan and Goldstein 1997; Taverna and Goldstein 2000).

We assess structural modularity and robustness indices, and an index of protein evolvability, for a data set of 167 mammalian proteins with empirically determined tertiary structures in order to look for an association between protein evolvability and either protein structural modularity or robustness. We find a positive association between our protein evolvability index and both structural modularity and robustness.

## Materials and Methods

Our experimental approach is to test whether proteins with high indices of evolvability are more structurally modular and/or robust than proteins with lower evolvability. Our data set consists of orthologous genes that code for proteins with solved tertiary structures. For each protein in the data set, we obtain measures of structural modularity, structural robustness, and evolvability.

The structural robustness index used here is contact density. In the context of relatively structured proteins (e.g., the data set used in this study), it can be assumed that the native fold is essential for function, so we can define robustness more specifically than we did in the Introduction. In this context, protein robustness is the ability for a protein sequence to maintain its native structure under mutational perturbation. Contact density is the average number of contacts an amino acid makes with other amino acids in the protein (England and Shakhnovich 2003). It has been shown to correlate with designability (England et al. 2003; England and Shakhnovich 2003; Bloom et al. 2006), which is the number of sequences that stably fold into a given structure and which is an important determinant of protein mutational robustness (Li et al. 1996; Bloom et al. 2005). Designability determines the rate at which stable folding becomes less likely as random mutations accumulate (Bloom et al. 2005; Wilke et al. 2005). High contact density implies many energetically favorable placements of strongly interacting amino acids, which relax energy constraints on the rest of the structure, thus allowing more sequences to fold into the structure (England and Shakhnovich 2003). We determine contact density using one of the standard methods (e.g., see Shakhnovich et al. 2005): we divide the trace (i.e., the sum of the elements on the main diagonal) of the square of the contact matrix by the number of residues in the protein structure. A contact matrix is calculated from the atomic coordinates of a protein database (PDB) structure file. We use the Euclidean distances between α-carbons to construct a distance matrix **D**. Using a threshold of 8 Å to define "contact," and excluding trivial contacts (defined as those between residues that are separated by fewer than two intervening residues in the sequence), we convert **D** to a Boolean contact matrix **C**, where 1 represents "contact" and 0 represents "no contact." Contact density is the trace of the square of **C**, divided by the number or residues in the protein: $\mathrm{Tr}(\mathbf{C}^2)/N$. Our specific methodological choices represent a compromise between the methods of H. Liao et al. (2005) who use α-carbons and a contact threshold of 9 Å, Shakhnovich et al. (2005) who use β-carbons and a threshold of 7.5 Å, and Bau et al. (2006) who use α-carbons and a threshold of 8 Å.

Since we are hoping to examine the relationship between modularity and evolvability, we would ideally measure protein modularity in a way that reflects the extent of evolutionary constraint. With our modularity measure, we aim to approximate the extent to which pleiotropic effects are restricted in the 3D space of the structure. A protein's independent units of evolutionary change (between which there are few pleiotropic effects) can be approximated through kinetic, thermodynamic, and/or functional modules (for a structural/folding perspective, see Copley et al. 2002 and for a functional perspective, Bhattacharyya et al. 2006). Here, we use structural modules as our approximation, which simply assumes that the constraining antagonistic pleiotropic effects of adaptive mutations are primarily due

to requirements for folding and stability. This is very likely the case for most of the proteins in this data set because they all have solved tertiary structures and so are biased to having rather rigid structures. For the reasons discussed below, we interpret helices (including α, $3_{10}$, and π helices) and β-sheet strands as structural modules. We can thus approximate the overall density of functional modules in a protein by simply dividing the number of helices and strands (defined according to the Dictionary of Protein Secondary Structure; Kabsch and Sander 1983) by the number of residues in the protein structure. We call this index "helix/strand density."

Although secondary structure elements are likely the smallest units that have some degree of evolutionary independence, it is surely the case that completely independent protein structural modules are generally larger than individual helices and strands. However, there are many reasons to believe that the majority of epistatic interactions between amino acids are highly localized within the 3D space of the protein structure. For one thing, the fundamental units of folding, function, and structure are clearly smaller than domains (del Sol and Carbonell 2007; Akiva et al. 2008; Laborde et al. 2008; Trifonov and Fenkel 2009). Evolutionarily independent motifs are also known to be very small (75% of them are between 10 and 40 residues; Su et al. 2005). Specific cases of particularly modular and evolvable protein domains also suggest that secondary structure elements are good descriptions of evolutionarily independent modules: The Duffy-binding-like domain is perhaps one of the most versatile and polymorphic protein domains in nature, and its ten semiconserved and evolutionarily independent sequence blocks have been shown to correspond near-perfectly with individual secondary structure elements (Howell et al. 2006). It was also shown that energetically independent modules have a mean size of 12 amino acids (Krishnan et al. 2007) and this corresponds well to the length of modules in our data set: while the mean length of individual secondary structure elements in our data set is 7.39 (standard deviation of the mean = 0.410), which is similar to previous calculations from empirical data (e.g., Sreerama et al. 1999), when mean module length is determined by dividing the full length of the protein structure (i.e., including both structured and unstructured sites), the mean module length is 13.8 (standard deviation of the mean = 0.504). Furthermore, Krishnan et al. demonstrate that energetically optimal modules correspond to single secondary structure elements until they reach about 30 amino acids in length (at which point energetically optimal modules of this length or longer are rare) (Krishnan et al. 2007). Finally, Emmert-Streib and Mushegian (2007) employ a method for domain identification that uses secondary structure elements as the fundamental units of structure, and they find that it performs equally well to more complicated analyses that include more detailed considerations of protein geometry and structure. This implies that secondary structure elements are the "main level at which protein domains attain their evolutionary optimal structural design" (Emmert-Streib and Mushegian 2007) and thus, that they offer a decomposition reflecting the protein's genuine epistatic architecture.

Another reason helices and strands are an appropriate choice is because the exact number of them within the protein structure can be easily and accurately ascertained from basic structural information. The small size of these structural modules also makes them more useful for constructing an informative modularity index because the number of them per protein structure is far more variable, and thus informative, than the number of larger entities, such as domains. For the above reasons, we think that helices and β-sheet strands provide the best description of protein modules that can be reliably determined for a large data set of proteins.

In the design of our structural modularity index, we choose to divide the number of modules by the total number of residues in the structure, as opposed to just the number of structured sites (i.e., those within helices or strands). Because all the proteins in our data set have solved tertiary structures, they are already biased to having a high proportion of structured sites, so our choice may be of little consequence. However, we make this choice because it is more conservative than the alternative. It is not clear that "unstructured" loop regions are free from all structural constraints on adaptation (e.g., Regad et al. 2010), so we choose to divide by the total number of residues to prevent any possibility of biasing our modularity measure to higher values in proteins with high proportions of unstructured sites. This type of bias could cause a problem for the interpretation of our results, since Ridout et al. (2010) find that the fraction of unstructured sites correlates with evolutionary rate. Though we do not find an association between the percentage of unstructured sites and evolvability in our data set, we design our modularity index so that we only risk reducing modularity measures in proteins with high fractions of unstructured sites. This assures that any bias in the index would only contribute to a negative correlation between structural modularity and evolvability, assuring that any positive correlation we detect would be a biologically meaningful signal.

Our evolvability index is an attempt to measure the extent to which positive selection, as compared to negative and neutral selection, determines protein sequence evolution. It measures the overall amount of adaptive evolution a protein experiences through its evolutionary history. It is a function of both the underlying constraints on adaptation and the extent to which the protein is exposed to forces of positive selection. Thus, it is more accurate to think of this as an index of realized evolvability. For example, even under strong positive selection, high structural and functional constraints can cause this index to be low, and in this sense,

it gauges the level of adaptive constraint. At the same time, however, this index will be low if there are low levels of positive selection—even when amino acid substitutions are unconstrained and free to evolve independently. For this reason, it is important to measure this index as an average over a large species tree because we aim to capture the long-term evolvability of the protein structure, in a range of contexts, rather than the particular selection pressures that may exist during the divergence of any two species. Because in this study we do measure the index as an average over a large species tree, we will not qualify it each time as an index for "realized evolvability." It will instead just be called an index for "evolvability."

Our evolvability index is the proportion of sites with evidence of being under positive selection multiplied by the average rate of adaptive evolution at these sites. Estimates for these numbers are obtained by analyzing the evolutionary history of each protein. For each of the proteins in the data set, a site model implemented by Phylogenetic Analysis by Maximum Likelihood (PAML) 3.15 codeml (Yang 1997, 2007) is used to analyze 25 mammalian orthologs mapped to a known species phylogeny (fig. 1). Parameters seqfile, outfile, and treefile were specified appropriately, and other parameters were set as follows: verbose = 1, seqtype = 1, CodonFreq = 2, aaDist = 0, model = 0, NSsites = 3, ncatG = 3, icode = 0, RateAncestor = 1, clock = 0, cleandata = 0, method = 0 (with all additional parameters being set to the codeml default settings, as described in the 2009 PAML version 4.3 user guide. From this analysis, we obtain the maximum likelihood estimates of the proportions of sites ($\rho_0$, $\rho_1$, and $\rho_2$) in each of three $\omega$ categories ($\omega_0$, $\omega_1$, and $\omega_2$), and the $\omega$ values themselves (where $\omega_0$ is constrained to be $<1$, $\omega_1$ is constrained to be $\leq 1$, and $\omega_2$ is left unconstrained) ($\omega = d_N/d_S$ = the ratio of the nonsynonymous substitution rate [$d_N$] to the synonymous substitution rate [$d_S$]). We define the proportion of sites with evidence of being under positive selection as $\rho_2$, and the rate of adaptive evolution at these sites as $\omega_2 - 1$, so our evolvability index is $\rho_2(\omega_2 - 1)$.

This evolvability index is importantly different from protein evolutionary rate indices used in many comparative studies (e.g., Bustamante et al. 2000; Fraser et al. 2002; Bloom and Adami 2003, 2004; Drummond et al. 2005; Herbeck and Wall 2005; Bloom et al. 2006; Chen and Dokholyan 2006; Lin et al. 2007). At least in theory (see qualification below), our index specifically measures the rate of substitutions that occur through positive selection.

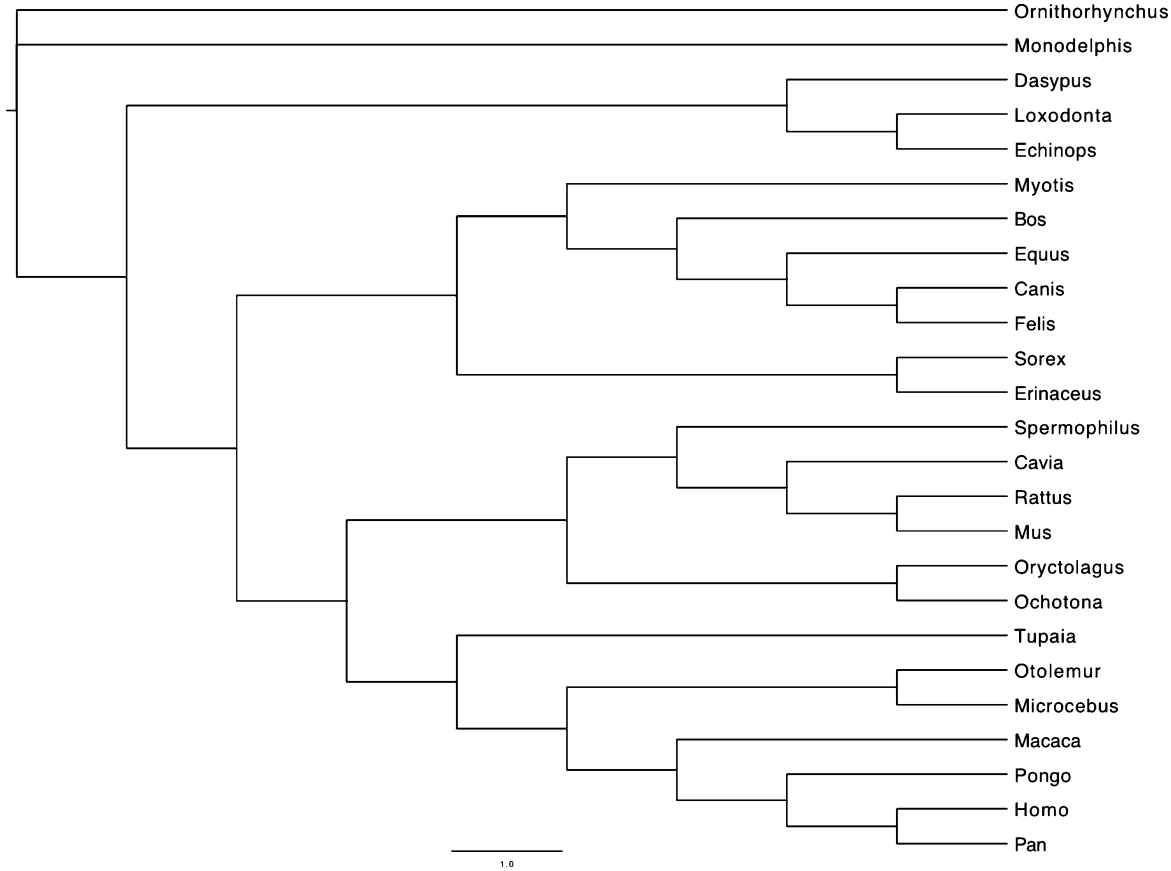


**Fig. 1.**—The species phylogeny for the 25 mammalian species that are represented in the OrthoMaM database as of February 2009 (Ranwez et al. 2007).

Conventional evolutionary rate indices take into account all types of substitutions and as a consequence (because neutral substitutions are so much more common than adaptive ones), they primarily reflect rates of neutral change. Because the ease with which a protein accommodates adaptive amino acid substitutions may not be directly related to the ease with which it accommodates neutral amino acid substitutions, if evolvability is defined as the ability to respond to positive selection (Wagner and Altenberg 1996; Pigliucci 2008; Wagner 2008), conventional evolutionary rate indices cannot serve as evolvability indices. Our evolvability index does, however, have one important weakness: For proteins without a class of sites under consistent and strong positive selection, it is possible that this index will overestimate the true level of adaptive evolution and be less negative than it should be, because when there are sites under significantly different levels of negative selection, it is possible for some sites to be identified as members of a third site class (i.e., an additional site class beyond those evolving primarily under neutral evolution and some specific level of negative selection) even when they do not experience significant positive selection. On the other hand, it should also be noted that an $\omega$ value below 1 (which would give an evolvability index below 0) does not imply that there are no sites under positive selection. It simply indicates that the average evolutionary rate across all branches of the phylogeny (fig. 1) is below 1 and thus dominated by negative selection. Thus, the assumption made here is that most proteins have at least some sites under positive selection on at least some branches of the mammalian tree and that this positive selection is significant enough that it is what generally defines the third class of sites with its distinct $\omega$ (as opposed to this being determined by the existence of two distinct rates of negative selection acting on different residues in the protein).

We obtain indices for structural modularity, structural robustness, and evolvability for 167 distinct proteins within the OrthoMaM database (Ranwez et al. 2007, accessed February 2009). This data set consists of all the proteins for which there is sufficient structural information to determine contact density and helix/strand density and for which orthologs of all 25 species are available. In this study, we limit our investigation to proteins from the same clade to eliminate potential confounding effects due to differences in phylogenetic structure between protein families from different groups. The data set is broken up into categories based on the broadest hierarchical Gene Ontology categories for molecular function (The Gene Ontology Consortium 2000), according to AmiGO version 1.7 (using the GO database release from 08 May 2010, Carbon et al. 2009). Within the data set, there are 155 proteins that have binding activity, 87 that have catalytic activity, 25 that have molecular transducer activity, 24 that have transcriptional regulatory activity, 16 that have enzyme regulatory activity, 6 that have transporter activity, 5 that have structural molecule activity, 1 that has electron carrier activity, and 5 with no known molecular function. The average values for contact density, helix/strand density, and the evolvability index are assessed for each of the eight molecular function subsets that have a sample size larger than 1. The data set is also broken up into two subsets according to whether the fraction of "structured" amino acids—that is, those that are part of a helix or strand—is relatively high or low. We also analyze the data set in three discrete subsets according to whether the secondary structure elements within the protein structure comprise only helices, only strands, or both.

To assess the nature of the relationship of protein evolvability to both structural modularity and robustness, we perform four tests. First, we perform a locally weighted polynomial fit to analyze the evolvability index as a function of structural modularity and robustness (with 0.5 of the data set used for each local fit). We carry out this analysis in R, using the "lowess" function. Second, each data set is divided into two equally sized groups according to the size of the evolvability index (dividing at the median value), and then Student's *t*-test, Welch's approximate *t*-test, and a Wilcoxon rank sum test are used to identify any significant difference between mean helix/strand density or contact density. We also compare the upper and lower third of the data set in a similar manner. Third, we perform Pearson's correlation and Spearman's rank correlation tests between the evolvability index and either contact density or modularity to look for any indication of an association between these two pairs of indices. Finally, our fourth test assess whether the variance in the evolvability index is significantly different for proteins with high versus low modularity or robustness: the data set is divided into two equally sized groups according to the size of either helix/strand density or contact density (dividing at the median value), and an *F* ratio test is performed between the two halves of the data set.

For the interpretation of our results, we rely on the assumption that different selection regime types are distributed approximately randomly across different protein fold types—that is, that the structural modularity and robustness of a protein does not significantly influence the selective forces it experiences. We test this assumption by looking for an association between protein functional importance and either helix/strand density or contact density. We measure functional importance by measuring the extent of negative selection acting on the protein, which is defined here as $\rho_0(1 - \omega_0)$.

We perform multiple regression to tease apart the separate influences of helix/strand density and contact density on the evolvability index. We divide the data set at the median value for the evolvability index and analyze the two halves separately. We determine the quadratic best-fit functions while constraining the functions to be equal to the median evolvability value at the lowest observed levels of helix/

strand density and contact density. We assess statistical significance of partial regression coefficients and compare the magnitude of standardized partial regression coefficients. We also perform a first-order polynomial multiple regression on the helix-only subset of the data set (since, as explained in Results, a linear fit was determined to be appropriate for this subset of the data).

To exclude possible confounding factors, we consider whether some additional protein variables co-correlate with our index of protein evolvability and either protein structural modularity or robustness. Gene compactness is the dominant factor determining evolutionary rate in mammals, and gene essentiality is among the factors of secondary importance (Liao et al. 2006). To determine whether it is necessary to control for gene compactness when assessing the relationship between the evolvability index and either helix/strand density or contact density, we test whether several compactness indices are significantly correlated with both the evolvability index and either helix/strand density or contact density. To determine whether it is necessary to control for gene essentiality, we assess whether there is a significant difference in the mean evolvability, structural modularity, or structural robustness index for essential versus nonessential proteins (i.e., those encoded by essential versus nonessential genes).

## Results

In this study, we test whether there is an association in proteins between structural modularity, structural robustness, and evolvability. We gauge protein structural robustness by assessing contact density—the average number of contacts an amino acid makes with other amino acids in the protein. We gauge structural modularity by assessing helix/strand density, which is the number of regular secondary structure elements divided by the number of residues in a protein structure. Unlike contact density, helix/strand density is not an established and well-studied index, so we test the basic assumption that underlies it: that the overall number of helices and strands correlates with the number of residues in a protein (if this were not the case, normalizing for protein size by dividing by the number of residues in the protein would over-correct for the influence of protein size). We find that there is a highly significant correlation between the number of residues and the number of helices and strands (fig. 2) and thus, that our normalization procedure is appropriate.

As described in detail in Materials and Methods, we obtain indices for modularity, robustness, and evolvability for 167 distinct mammalian proteins, each with orthologs from 25 species. To assess whether protein structural modularity and robustness have an influence on protein evolvability, we test whether there is a positive association between the evolvability index and either structural modularity or robustness. The evolvability index is plotted as a function of both the structural robustness index (contact density) and the structural modularity index (helix/strand density) (fig. 3*A*, and *B*). The contact densities of the proteins in the data set have a mean value of 5.1 and a standard deviation of 1.0, the helix/strand densities have a mean of 0.082, and a standard deviation of 0.023, and the evolvability indices have a mean of −0.0095 and a standard deviation of 0.066.

The relationship between contact density and the evolvability index reveals two interesting and significant patterns. First, a general positive association between these two indices is apparent when we perform a locally weighted polynomial fit that provides a sliding window analysis of the relationship between the evolvability index and contact density (fig. 3*C*). While it seems that the average evolvability index generally increases with increasing contact density, this analysis also reveals that the relationship is not simple or linear. This general positive association between contact density and the evolvability index is also reflected in hypothesis test results: When the sample of proteins is divided into two equally sized groups according to their evolvability index (less-than-median vs. greater-than-median), we find that the mean contact density of the group with relatively high evolvability indices (5.30) is significantly greater than the mean contact density of the group with relatively low evolvability indices (4.94) ($P = 0.0101$ for Student's *t*-test, $P = 0.0100$ for Welch's approximate *t*-test, and $P = 0.0125$ for Wilcoxon rank sum test with continuity correction, all one-tailed) (fig. 3*A*). Much of the data set is highly clustered with respect to the evolvability index, and very small differences can be of questionable biological relevance even when they are statistically significant. We therefore also compared the highest and lowest thirds of the data set with respect to the evolvability index and found that the mean contact density of these two smaller groups still differs significantly. Further evidence for the positive association between contact density and the evolvability index is that these two indices have a borderline significant rank correlation ($P = 0.0682$ for Spearman's rank correlation test, one-tailed), though they do not have a significant linear correlation ($P = 0.391$ for Pearson's correlation test, one-tailed).

The above patterns imply that proteins with higher evolvability indices are generally more designable and robust than proteins with lower evolvability indices. These patterns also prove to be even more pronounced when we look only at proteins that contain helices but no strands (fig. 4*A*). Analyzing this subset of data independently, we find a significant rank correlation between the indices but not a significant linear correlation ($P = 0.119$ for Pearson's correlation test, one-tailed; $P = 0.0162$ for Spearman's rank correlation test, one-tailed). Also, when this subset of data is divided in half according to the median evolvability index, we find that the difference between the mean contact density is significant

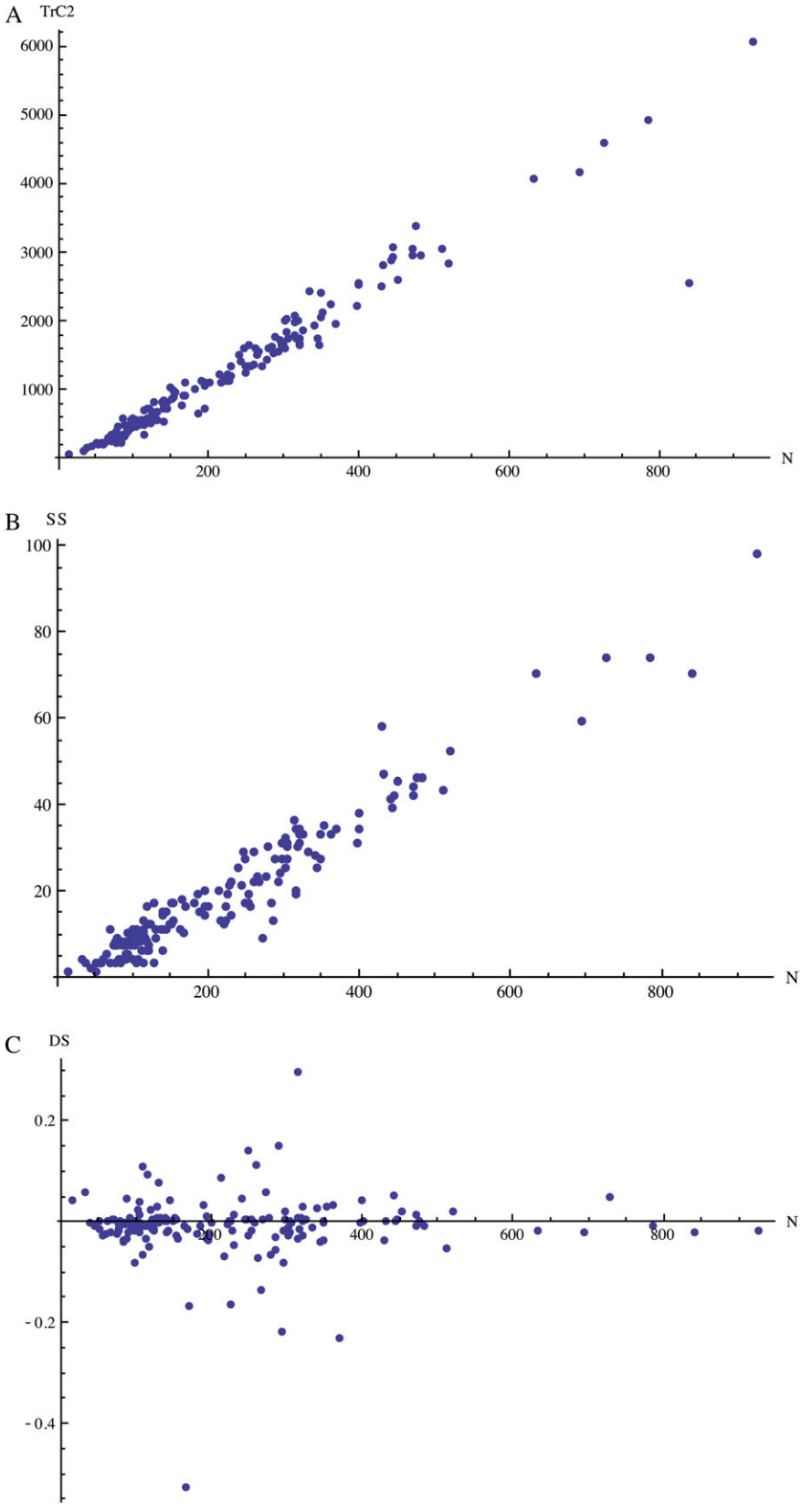


**Fig. 2.**—(*A*) The trace of the square of the contact matrix "TrC2" as a function of the number of residues in the protein structure "N." Pearson correlation coefficient = 0.970. (*B*) The total number of helices and strands "SS" in a protein structure as a function of the number of amino acids in the protein structure "N." Pearson correlation coefficient = 0.966. (*C*) The evolvability index "DS" as a function of the number of amino acids in the protein structure "N."

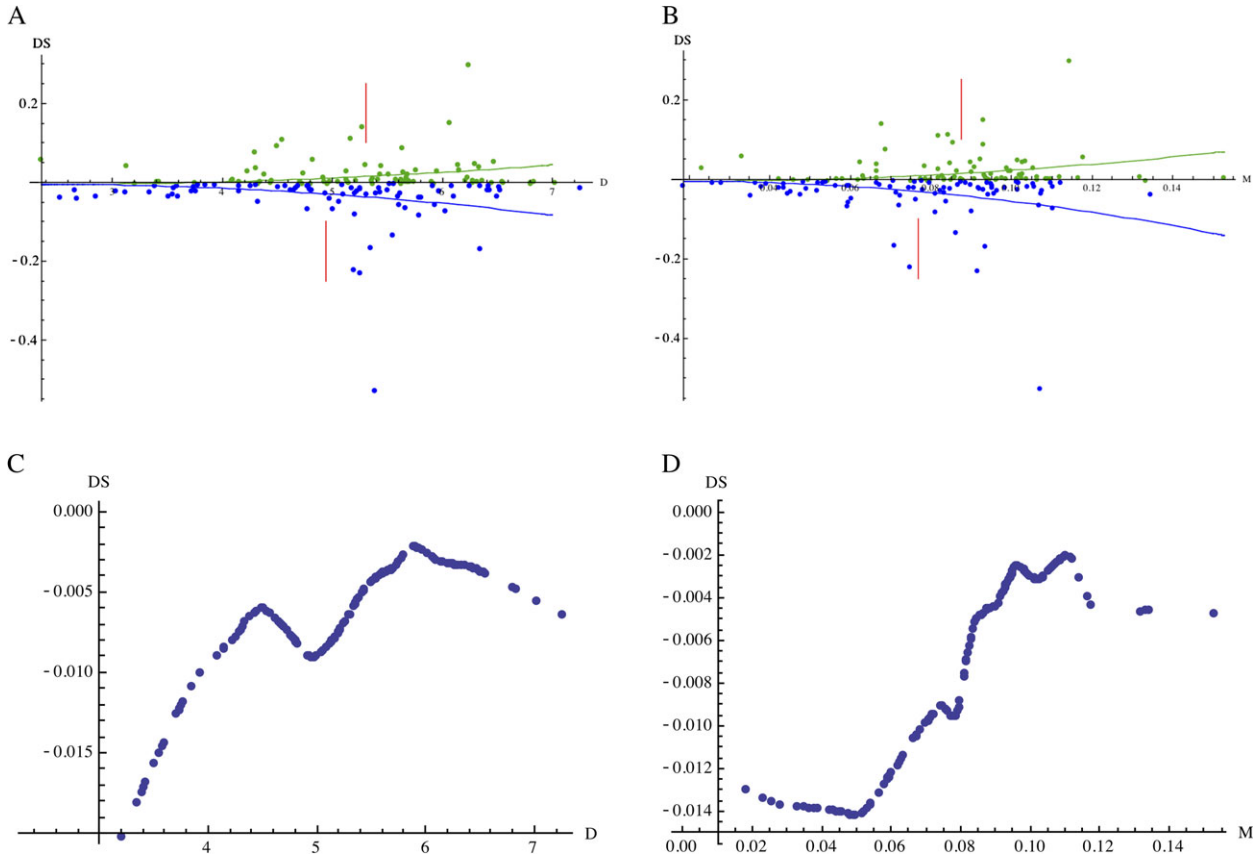


**Fig. 3.**—(*A*) The evolvability index "DS" as a function of the structural robustness index (contact density) "D." Spearman's rank correlation coefficient = 0.116, one-tailed $P = 0.0682$. (*B*) The evolvability index "DS" as a function of the structural modularity index (helix/strand density) "M." Spearman's rank correlation coefficient = 0.164, one-tailed $P = 0.0169$. (*A*, *B*) The color of the data points indicates whether they are part of the upper or lower half of the data set with respect to "DS" divided at the median. The mean "D" or "M" of the light green data points is indicated by the upper line, and the mean "D" or "M" of the dark blue data points is indicated by the lower line. Curves are best-fit parabolic functions without a constant basis and constraining "DS" to be equal to the median "DS" value for the lowest observed "D" or "M" value. Both fits are highly significant ($P << 0.0001$ according to analysis of variance *F* statistic). (*C*, *D*) Sliding-window analysis (i.e., locally weighted polynomial regression) of the mean evolvability index "DS" as a function of contact density "D" (C) or helix/strand density "M" (D). The proportion of the data set used to fit each local polynomial is 0.5.

($P = 0.0394$ for Student's *t*-test, one-tailed; $P = 0.0395$ for Welch's approximate *t*-test, one-tailed; $P = 0.0366$ for Wilcoxon rank sum test with continuity correction, one-tailed) and greater than it is for other subsets of the data (fig. 5*A*)

In addition to the above evidence for a positive association between the evolvability and structural robustness indices, we observe a second pattern between these indices: high contact density seems to be associated with greater variance in the evolvability index across different proteins (fig. 6*A*). Indeed, when the data set is divided into two equally sized groups according to contact density (dividing at the median), the variance in the evolvability index is significantly greater for proteins with higher contact density than for those with lower contact density (0.00164 vs. 0.00718) ($P << 0.0001$ for *F* ratio test of null hypothesis that ratio between the variances is 1). Furthermore, this difference in variance is not dependent on the outlying data points: if the two most outlying data points with respect to the evolvability index are removed from both halves of the data set, there is still a significant difference between the variances of the two halves. Thus, we observe an increase in the variance of the evolvability index as contact density increases.

In order to analyze the relationship between structural modularity and evolvability in proteins, we perform the same tests as above, but this time for helix/strand density. We perform a local fit on the full data set to analyze the relationship between the evolvability index and helix/strand density (fig. 3*D*). As with contact density, this analysis reveals that the evolvability index generally increases with increasing helix/strand density. However, it also shows that the relationship between these two indices is not necessarily a simple linear one.

We find that the mean helix/strand density of proteins with relatively high evolvability indices (0.0876) is significantly greater than the mean helix/strand density of proteins

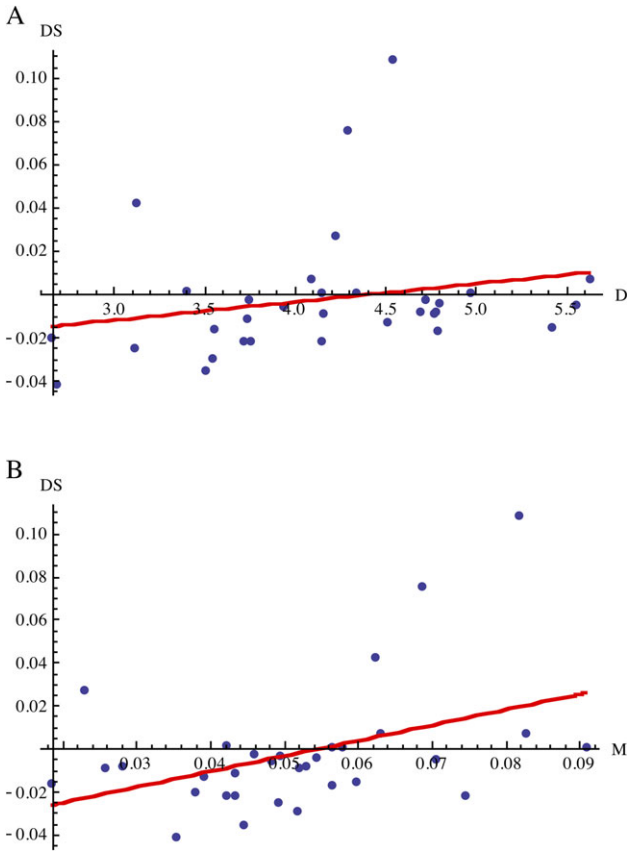


**Fig. 4.**—Correlation and rank correlation analysis for a subset of the data set, consisting of proteins that contain helices, but no strands. (*A*) The evolvability index "DS" as a function of the structural robustness index (contact density) "D." The best-fit line is shown in red but is not statistically significant (Pearson's correlation coefficient = 0.215, one-tailed $P = 0.118$), but the rank correlation between DS and D is significant (Spearman's rank correlation coefficient = 0.380, one-tailed $P = 0.0162$). (*B*) The evolvability index "DS" as a function of the structural modularity index (helix/strand density) "M." The best-fit line is shown in red and is statistically significant (Pearson's correlation coefficient = 0.414, one-tailed $P = 0.0092$), and the correlation between "M" and "DS" is statistically significant even after correcting for the relationship between "D" and "DS". The rank correlation between the variables is also significant (Spearman's rank coefficient = 0.426, one-tailed $P = 0.0748$).

with relatively low evolvability indices (0.0768) ($P = 0.00135$ for Student's *t*-test, one-tailed; $P = 0.00135$ for Welch's approximate *t*-test, one-tailed; $P = 0.00324$ for Wilcoxon rank sum test with continuity correction, one-tailed) (fig. 3*B*). The difference in mean helix/strand density between the highest and lowest thirds of the data set with respect to the evolvability index is also significant. These hypothesis tests confirm the generally positive association between secondary structure density and the evolvability index observed in the local fit. Furthermore, while helix/strand density does not significantly correlate (i.e., linearly) with the evolvability index ($P = 0.375$ for Pearson's correlation test, one-tailed), it does significantly rank correlate with the evolvability index ($P = 0.0169$ for Spearman's rank correlation test, one-tailed).

The evidence for a positive association between secondary structure density and the evolvability index is stronger when we analyze the subset of data consisting of helix-only proteins (fig. 4*B*). In this case, we find a significant linear correlation as well as rank correlation ($P = 0.0092$ for Pearson's correlation test, one-tailed; $P = 0.00748$ for Spearman's rank correlation test, one-tailed). Also, when this subset of the data is divided in half at the median evolvability index, the difference between the mean secondary structure density for the two halves of the data set is significant ($P = 0.00400$ for Student's *t*-test, one-tailed; $P = 0.00409$ for Welch's approximate *t*-test, one-tailed; $P = 0.00332$ for Wilcoxon rank sum test with continuity correction, one-tailed) and greater than the difference between the means for other subsets of the data (fig. 5*B*).

Finally, as in the case of structural robustness, the evolvability indices of proteins with relatively high structural modularity are significantly more variable than those of proteins with relatively low structural modularity (0.00657 as compared with 0.00224; $P << 0.0001$ for *F* ratio test of null hypothesis that ratio between the variances is 1) (fig. 6*B*). This result holds regardless of whether outlying data points are included or not (the difference in variance is significant even if the two most outlying data points with respect to the evolvability index are removed from both halves of the data set). Together with the corresponding results for contact density, this implies that higher structural modularity and robustness are associated with greater variance in realized protein evolvability.

Because our indices for structural modularity and structural robustness correlate with one another to some extent (fig. 7), the above results on their own do not clarify whether either of these indices have independent effects on protein evolvability. We therefore perform multiple regression to tease apart the separate influences of helix/strand density and contact density on the evolvability index. The full data set is divided at the median evolvability index, and the two halves are analyzed separately. Quadratic fits to both halves of the data set are highly significant (analysis of variance $P << 0.0001$). The estimates of the individual partial regression coefficients—the parameters that describe how helix/strand density and contact density independently influence the evolvability index—were not significantly different from 0 in either case (Student's *t*-test). Thus, the relative statistical significance of the partial regression coefficients cannot be used to exclude either helix/strand density or contact density as a possible independent predictor of the evolvability index. We find that, for both halves of the data set, the standardized partial regression coefficient for helix/strand density is nearly 100 times greater in magnitude than the standardized partial regression coefficient for contact density (regardless of the order in which the two variables

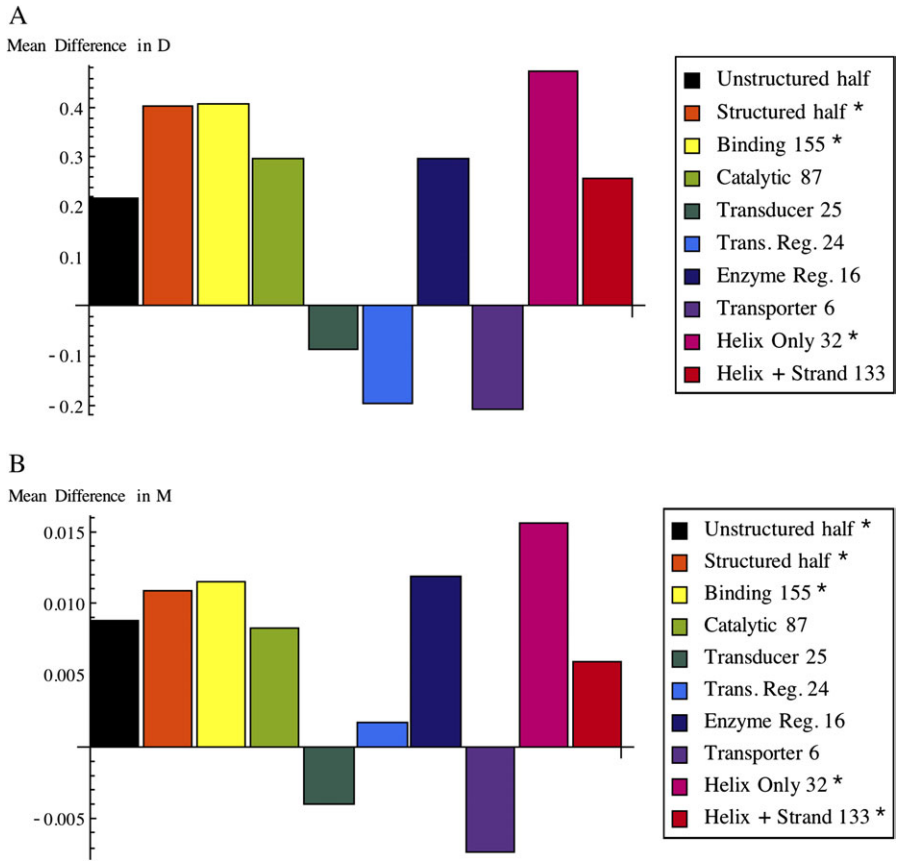


**Fig. 5.**—Subsets of the data set (those with a sample size greater than 5) are analyzed independently. The data subsets are each divided into upper and lower halves with respect to their evolvability indices (divided at the median), and then the difference between (*A*) the mean contact densities "D" and (*B*) the mean helix/strand densities "M" for the upper and lower halves of the data set are assessed. In the legend, numbers indicate sample size and asterisks indicate significance of the difference between the means according to one-tailed Student's *t*-test with a significance cutoff $P = 0.05$.

are added to the model), however, we cannot conclude anything from this because the partial regression coefficients are not significantly different from 0.

As stated above, the helix-only proteins show a stronger pattern, and for this subset of the data, the relationships between helix/strand density, contact density, and evolvability can all be meaningfully approximated as linear (fig. 4). We therefore perform multiple regression on this subset of data that comprises helix-only proteins using a first-order polynomial fit. We find that helix/strand density is a significant contributor to the variation in the evolvability index even after controlling for the influence of contact density (the standardized partial regression coefficient for helix/strand density is 0.387 and $P = 0.0405$), whereas this is not the case the other way around (the standardized partial regression coefficient for contact density is 0.0763 and $P = 0.675$). Both the magnitude of the standardized partial regression coefficients and the difference in whether they are significant demonstrate that helix/strand density is more important than contact density in determining the value of the protein evolvability index.

In addition to analyzing helix-only proteins in isolation, we examined several other subsets of the data independently. Specifically, we separately analyzed the subset of proteins that contains strands and no helices, and the subset that contains a mixture of helices and strands. We also divided up the data set for separate analysis according to the molecular function of the proteins and according to whether they are "structured" or "unstructured" (see Materials and Methods). We find that the mean helix/strand density and contact density do differ between these various categories (fig. 8). We also find that contact density is negatively correlated with the fraction of unstructured sites and that structural modularity is positively correlated with the fraction of unstructured sites but that the evolvability index is not associated positively or negatively with the fraction of unstructured sites. We find no major incongruencies among these data subsets in regard to the relationship they reveal between the indices for modularity, robustness, and evolvability (figs. 5 and 9). However, these different subsets reveal the relationships between the indices to varying extents. As mentioned above, the helix-only category of proteins reveals

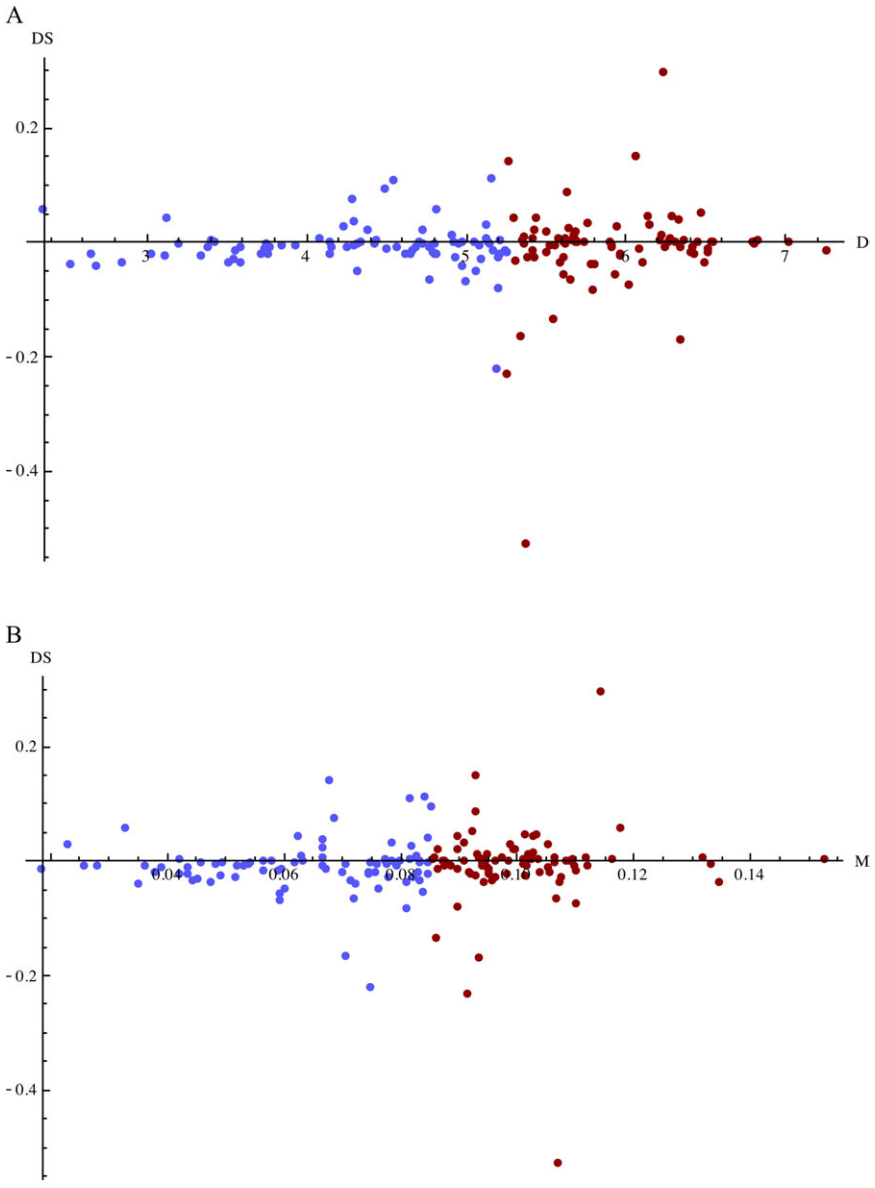


**Fig. 6.**—(*A*) The evolvability index "DS" as a function of contact density "D." (*B*) The evolvability index "DS" as a function of helix/strand density "M." (*A*, *B*) The color of the data points indicates whether they are part of the upper or lower half of the data set with respect to "D" or "M", divided at the median. The variance along the *y* axis of the red data points is significantly larger than the variance along the *y* axis of the blue data points.

much stronger positive associations between the evolvability index and both contact density and helix/strand density (figs. 4 and 5). Conversely, the limitations of helix/strand density and contact density to act as indicators of the level of adaptive constraint is reflected in the fact that these patterns are considerably less pronounced for classes of proteins known to be highly unstructured (fig. 5) (Wright and Dyson 1999; Garza et al. 2009) and for the less structured half of the data set (figs. 5 and 10). This implies that these structural indices fail to capture the relevant constraints on adaptation for unstructured proteins, as expected.

## Testing for Potential Confounding Factors

To index evolvability in proteins, we measure the amount of adaptive evolution a protein experiences. As mentioned above, in using this index, we are assuming that high levels of evolution through positive selection can be attributed at least partially to low constraints on adaptation (i.e., high evolvability) as opposed to only high positive selection

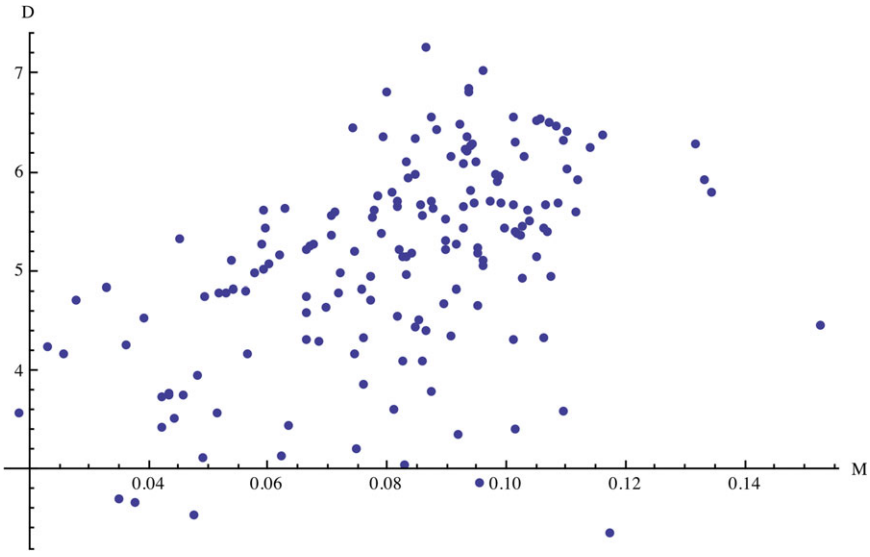


**Fig. 7.**—Helix/strand density "M" as a function of contact density "D."

pressure, and that the structural modularity or robustness of a protein does not significantly influence the selective forces it experiences. It is important that these assumptions are true because a confounding cause of our results would be that proteins with high structural modularity or robustness for some reason experience preferentially higher positive selection pressure. To verify that this is not the case, we look for whether functional importance is associated with structural modularity or robustness. If functionally important proteins—which we define to be those under strong negative selection—are generally more modular and robust than less important proteins, we would have to consider the possibility that our indices for structural modularity and robustness only correlate with evolvability due to recruitment of modular and robust folds into important functional roles or through gradual selection for increased modularity or robustness in important proteins (though this latter possibility is unlikely for reasons discussed below). However, we do not find any association between the index for functional importance and either helix/strand density or contact density (supplementary fig. S1, Supplementary Material online), so we reject these possible confounding causes of our results.

According to a recent study by Ridout et al. (2010), unstructured sites (i.e., those which are not part of a regular secondary structure element) are more likely to have high ω values. This poses a possible alternative explanation for our observed association between structural modularity and evolvability indices (figs. 3*B*, 3*D*, and 4*B*): i.e., that it is just a trivial consequence of there being a greater proportion of unstructured sites in highly modular proteins. This is especially plausible since we also happen to find that proteins with higher proportions of unstructured sites (defined here as those not within a helix or strand) tend to have higher helix/strand density (supplementary fig. S2, Supplementary Material online). However, we rule out this alternative interpretation because our evolvability index shows no association with the proportion of unstructured sites (supplementary fig. S3, Supplementary Material online).

Our index of structural robustness—contact density—has been previously shown to correlate with protein length (Lipman et al. 2002; Bloom et al. 2006), and we find this correlation in our data also (supplementary fig. S4, Supplementary Material online). To rule out the possibility that the association between contact density and the evolvability index (figs. 3*A*, 3*C*, and 4*A*) is caused by a co-correlation of both indices to protein length, we test for whether there is any relationship between evolvability and protein length. We find no significant correlation between these two variables (fig. 2*C*). Furthermore, when we divide the data set into two groups (one comprising those with less-than-median protein length and the other comprising those with greater-than-median protein length), we find no significant difference in the mean evolvability indices of these two groups.

Liao et al. (2006) demonstrate that gene compactness and gene essentiality are both important determinants of the overall rate of mammalian protein evolution. To determine whether it is necessary to control for gene compactness when examining the relationships between protein structural modularity, robustness, and evolvability, we test whether gene compactness indices co-correlate with the evolvability index and either protein structural modularity or robustness (supplementary figs. S5, S6, and S7, Supplementary Material online). We found no co-correlations and we find only two significant negative correlations among all

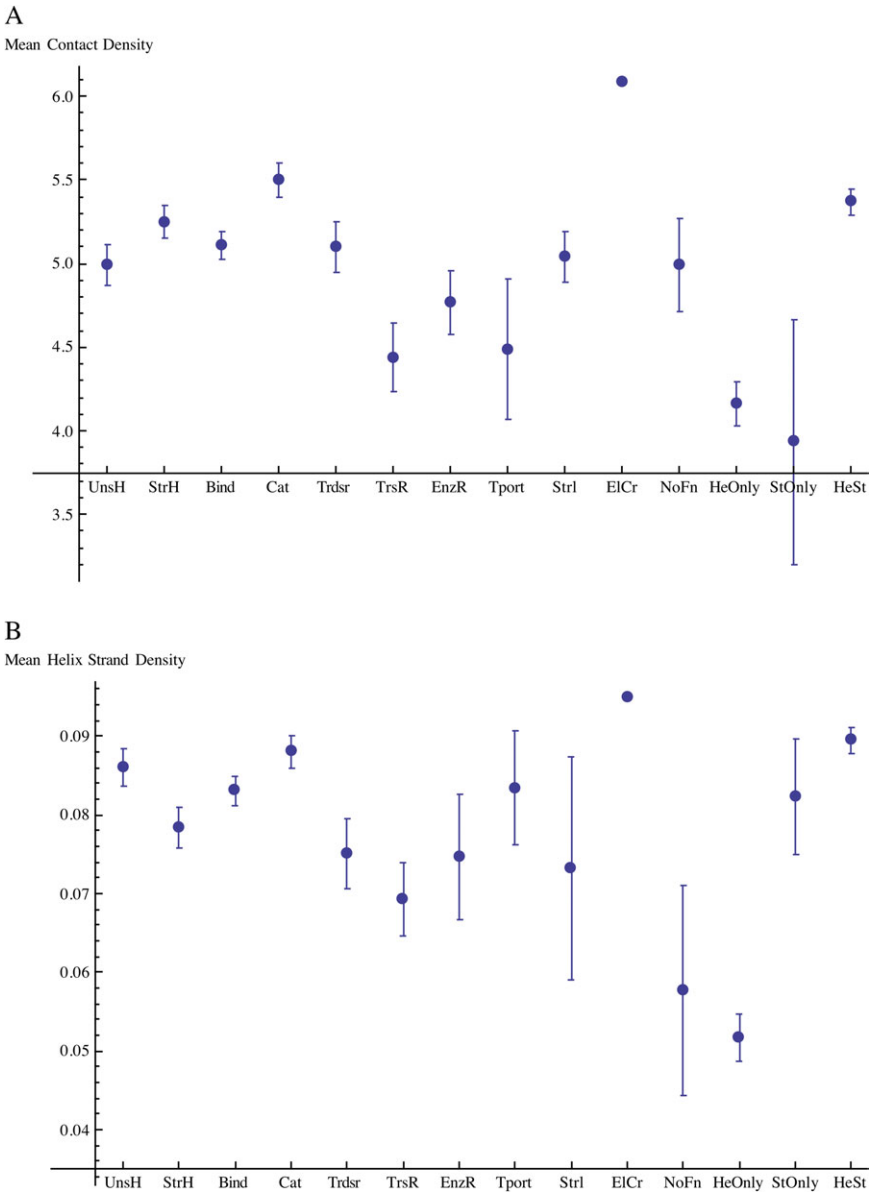


**Fig. 8.**—(*A*) The mean structural robustness index (contact density) for proteins of different fold and functional categories. (*B*) The mean structural modularity index (helix/strand density) for proteins of different fold and functional categories. UnsH, relatively unstructured half of data set; StrH, relatively structured half of data set; Bind, binding activity; Cat, catalytic activity; Trdsr, molecular transducer activity; TrsR, transcription regulator activity; EnzR, enzyme regulator activity; Tport, transport protein activity; Strl, structural molecule activity; ElCr, electron carrier activity; NoFunc, no molecular function specified; HeOnly, secondary structure elements consist of helices only; StOnly, secondary structure elements consist of strands only; HeSt, secondary structure elements consist of both helices and strands. Error bars show the standard error of the sample mean and are included where the sample size for the category is above 1.

the tests we perform—between CDS length and contact density and between CDS length and helix/strand density (before correcting for multiple tests, $P << 0.001$ and 0.047, respectively). Because CDS length does not also negatively correlate with the evolvability index, we conclude that CDS length cannot be responsible for the observed associations between the evolvability index and structural modularity and robustness. To determine whether it is necessary to control for gene essentiality, we assess whether there is a significant difference in helix/strand density, contact density, or the evolvability index in proteins corresponding to essential versus nonessential genes (supplementary fig. S8, Supplementary Material online). We find no significant differences among these comparisons (with the significance cutoff set to $P = 0.05$ before correcting for multiple tests). Therefore, we conclude that gene essentiality is not

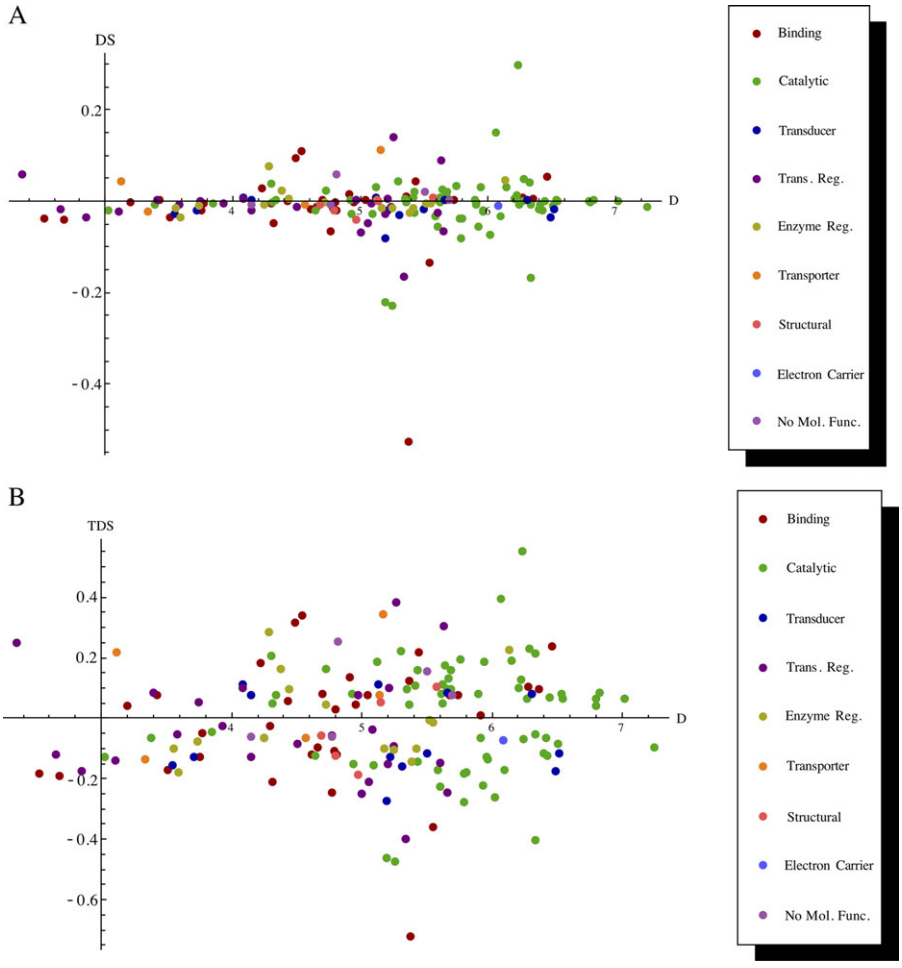


**Fig. 9.**—(*A*, *B*) Protein structural robustness "D" with proteins categorized by molecular function. The legend lists molecular functions in the order of their frequency in the data set, starting at the top with the most frequent. Where a protein has more than one molecular function, it is specified as the least frequent one. (*A*) The evolvability index "DS" as a function of the structural robustness index (contact density) "D." (*B*) Transformed evolvability index "TDS" as a function of "D". Transformation of "DS," by subtracting the median value and then taking the square root, allows for better visualization of the data. (*C*, *D*) Protein structural modularity "M" with proteins categorized by molecular function. The legend lists molecular functions in the order of their frequency in the data set, starting at the top with the most frequent. Where a protein has more than one molecular function, it is specified as the least frequent one. (*C*) The evolvability index "DS" as a function of the structural modularity index (helix/strand density) "M." (*D*) Transformed evolvability index "TDS" as a function of M. Transformation of "DS," by subtracting the median value and then taking the square root, allows for better visualization of the data.

likely to be a confounding factor in our analysis of the relationship between structural modularity, structural robustness, and evolvability.

## Discussion

From a theoretical standpoint, a system must be robust to be evolvable by natural selection. And yet, it remains unclear whether the ubiquity of robustness in nature can be explained by selection for evolvability or whether it has evolved for the sake of buffering mutational and/or environmental noise (Hartl and Taubes 1996; Wagner et al. 1997; Ancel and Fontana 2000; Meiklejohn and Hartl 2002; de Visser et al. 2003; Wagner 2005). Modularity is another characteristic of biological systems with obscure origins, and it is thought to contribute to robustness and evolvability (Wagner et al. 2007). Investigation into the origins of modularity and robustness is stymied by the fact that there is scant empirical evidence that they are biologically significant determinants of evolvability, probably because defining and measuring modularity and robustness in real biological systems remains problematic. Here, we use one established index of protein structural robustness (contact density as a measure of designability) and another index of our own design (helix/strand density as a measure of structural modularity) to test whether robustness and modularity are associated with evolvability in proteins. Prior to this study, we knew little

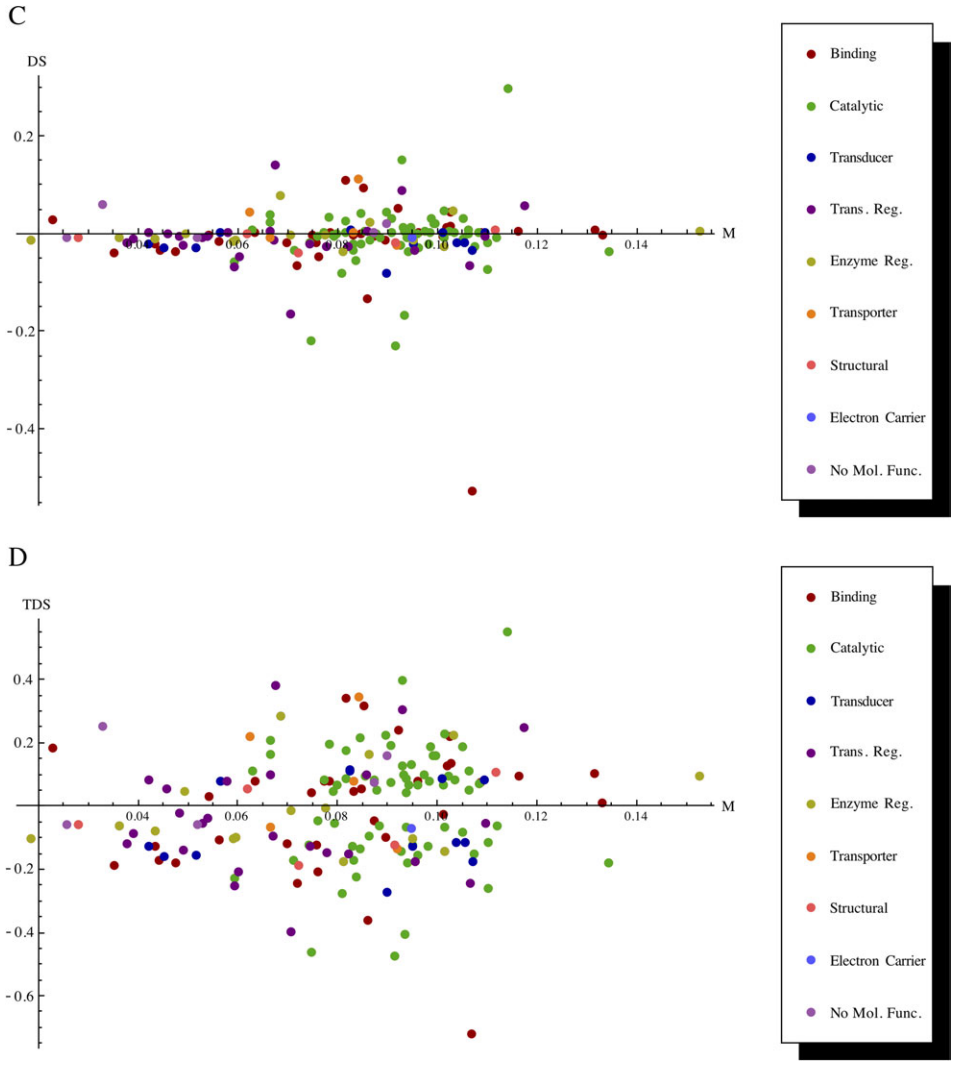


**Fig. 9.**—(Continued)

about the distribution of helix/strand density across different proteins, but previous work had already established that contact density is a determinant of protein family size (Shakhnovich et al. 2005), sequence diversity (Hartling and Kim 2008), functional diversity (Ferrada and Wagner 2008), and overall evolutionary rate ($d_N$) in yeast (Bloom et al. 2006). These studies provide some indication that contact density contributes to reduced constraints and possibly evolvability. However, Bloom et al. (2006) could not fully disentangle the effects of contact density and protein length on $d_N$, so it is possible that contact density only correlates with $d_N$ through co-correlation with protein length or some other unmeasured factor (such as modularity). Furthermore, these studies do not infer evolvability by measuring the amount of evolutionary change brought about through positive selection, as we do here. Instead they use protein family size, $d_N$, or functional or sequence diversity, which are all influenced by more factors than the two which contribute to our evolvability index (i.e., the extent of constraints on adaptation and positive selection strength).

## Protein Structural Modularity and Robustness and Their Effects on Protein Evolvability

Here, we address whether structural modularity and robustness contribute to evolvability in proteins. We hypothesize that high values for either structural modularity or robustness should reflect low structural constraints and since these likely represent the dominant constraints in structured proteins, high evolvability. Therefore, if modularity and robustness confer evolvability in proteins—assuming different selection regimes are distributed approximately randomly among different protein folds—we expect to find a positive association between our index for protein evolvability and

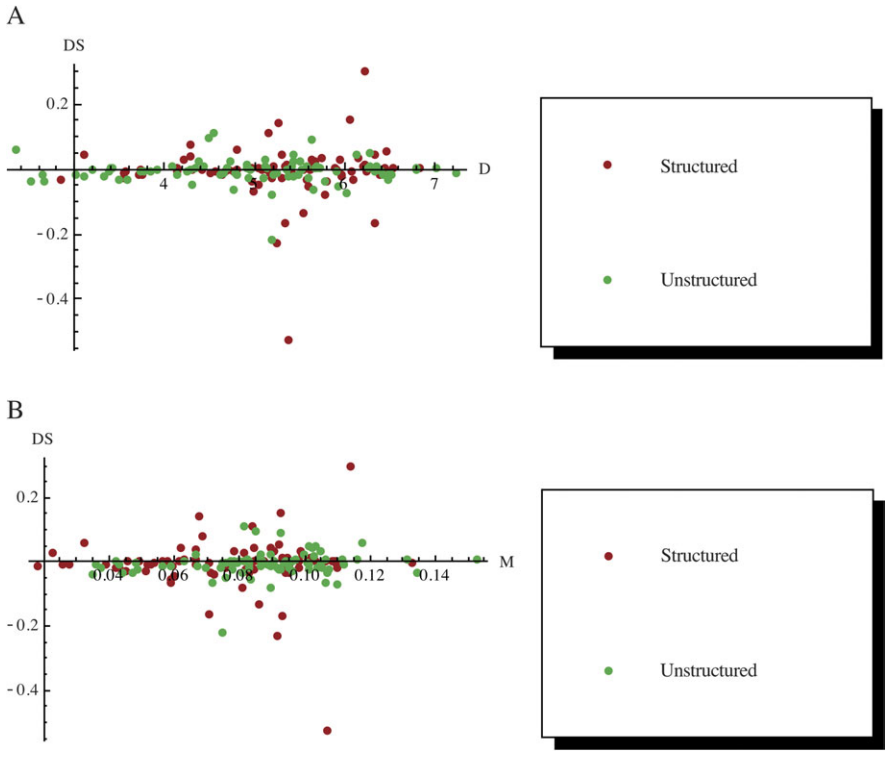


**Fig. 10.**—The evolvability index "DS" as a function of (*A*) the protein structural robustness index (contact density) "D" and (*B*) the protein structural modularity index (helix/strand density) "M." The data set was divided into two equally sized groups according to a protein's proportion of "structured" sites—defined as those that are part of a helix or strand.

both helix/strand density and contact density. Indeed, this is what we find. Specifically, we find that 1) a sliding-window analysis, which fits distinct polynomials to local subsets of the data, reveals a generally positive association both between contact density and the evolvability index and between helix/strand density and the evolvability index (figs. 3*C*,*D*) and that this pattern is stronger when we consider only the proteins containing helices and no strands (fig. 4); 2) the evolvability index is significantly rank correlated with the modularity index and, to a lesser extent, the robustness index (fig. 3) and that this is especially the case for helix-only proteins (fig. 4); and 3) multiple regression analysis of helix-only proteins demonstrates that the correlation between the modularity and evolvability indices is independent of contact density and that this is not the case the other way around—implying that helix/strand density is more important than contact density in determining the value of the evolvability index and that the apparent association between contact density and evolvability in proteins may be driven by co-correlation of structural modularity to both contact density and evolvability.

We rule out the possibility that differences in evolvability are due to differences in selection regime when we fail to find an association between structural modularity or robustness and protein functional importance (supplementary fig. S1, Supplementary Material online). This finding means that we can exclude two possible alternative interpretations. The first is that, in the long term, robust protein folds—being more evolvable—end up being recruited into functional roles which demand high levels of evolvability because they are good at tolerating shifting selection pressures (in other words, the possibility that highly robust proteins are predisposed to biological roles where adaptive changes are frequent and that protein robustness persists through association with these adaptive changes). This would constitute a mechanism of fold selection for evolvability (i.e., selection for the most evolvable fold, out of multiple distinct folds that can perform the same function) (Taverna and Goldstein 2000; England et al. 2003). The second alternative interpretation is that strong positive selection, which would be reflected as high levels of adaptive evolution, causes proteins to gradually evolve greater robustness. From a theoretical standpoint, this interpretation is unwieldy to begin with because contact density and helix/strand density, being inherent features of the protein structure, cannot evolve efficiently through point mutations because distinct protein structures are separated in sequence space by vast distances composed almost entirely of unfoldable sequences (i.e., there is no shape space covering) (Babajide et al. 2001). Hence, one of the basic requirements for adaptive evolution—that the trait can change in a quasi-gradual way—is not fulfilled by either helix/strand density or contact density.

We find that higher structural modularity and robustness are associated with greater variance between proteins in the evolvability index (fig. 6). We think this is most likely due to the fact that the evolvability index is the product of two variables, at least one of which (i.e., the proportion of sites under positive selection) is very likely to be binomially distributed and thus have increasing variance with increasing magnitude. This is because the proportion of sites under positive selection may simply be a sum of multiple independent Bernouli trials that each determine whether or not a given site is under positive selection), divided by the total number of sites. Accordingly, since the number of successful outcomes of any Bernouli trial is binomially distributed, if we specify $P$ as the probability that any given site is under positive selection and $n$ as the total number of sites in the protein, we can describe this proportion of sites under positive selection as a binomially distributed variable with an expected value $P$, and a variance $np(1 - P)/n^2$. Thus, we expect that both the mean and variance will increase with $P$ so long as $P$ is less than 0.5—which is certainly the case for the proteins in our data set.

## The Relationship between Protein Structural Modularity and Robustness

There are some other minor conclusions that can be drawn from this work. By quantifying both protein structural modularity and robustness, we have the opportunity to address how these two variables relate to one another. The exact relationship between them has not been thoroughly investigated in real proteins. All that is known is that, for lattice models, mutationally robust "prototype" sequences are characterized by an overrepresentation of special sequence motifs that fold in a context-insensitive manner—reminiscent of "folding modules" (Cui et al. 2002). Also, Li et al. (2007) show that modular "stabilizing fragments" can be recombined to create highly robust chimeric proteins. Lastly, for approximately factorizable networks, theory shows that the mean clustering coefficient (which is an index for modularity) is determined by the heterogeneity and density of the network. Though it is not clear whether proteins represent approximately factorizable networks, single domain proteins do seem to fit their general profile. Network density is very related to contact density when it is applied to amino acid structural networks (Dong and Horvath 2007), so this could be what causes the correlation we find between contact density and helix/strand density. However, because we find that contact density does not tightly correlate with helix/strand density (fig. 7) and that helix/strand density has an effect on the evolvability index that is independent of the effect of contact density, we conclude that structural modularity and structural robustness—at least as indexed here—describe somewhat different information. However, because they are also clearly intertwined, our results emphasize the importance including considerations of protein structural modularity in studies involving contact density and the value of developing methods to quantify modularity in real biological systems.

## Robustness of Unstructured Proteins

It is important to note that our indices for both modularity and robustness are structural and that structural constraints are only good approximations of the overall constraints on adaptation where structure is essential for function. While this is true for many proteins, there are some important exceptions. For example, many transcription factor proteins only require structural stability at a small fraction of their amino acids (Garza et al. 2009) and disordered regions often have important functional roles and conserved sequences (Marisco et al. 2010). Moreover, it has been hypothesized that proteins without a rigid structure achieve high robustness of function (despite essentially zero structural robustness) (Brown et al. 2002), flexibility of function for transient and specific interactions (Singh et al. 2007), and the ability to evolve through promiscuous functions (Wroe et al. 2007)—all of which contribute to higher evolvability. Our results indicate that structural constraints do not capture the relevant constraints on adaptation for some classes of proteins in our data set—specifically, those which are relatively unstructured (figs. 5, 8, and 10). Therefore, our results support the idea that, for some proteins, proper function is not directly dependent on structural stability, and in turn, that protein functionality cannot always be approximated through measures of structural stability. This is significant in light of the common assumption within the field of structural biology that structure equals function. However, because the great majority of proteins with solved structures do rely on an ordered structure to perform their functions, we did not think that these exceptions would cause enough of a problem to warrant their exclusion from our data set.

## Future Research about the Determinants of Protein Evolutionary Rate

There has been considerable research in the past several years aiming to identify the important determinants of protein evolutionary rate ($d_N$ or $d_N/d_S$). For the reasons stated above, we believe that our evolvability index is fundamentally different from these measures of predominantly neutral evolutionary change. Furthermore, in these studies about the determinants of evolutionary rate, $d_N$ or $d_N/d_S$ is generally inferred from a comparison of only two species, whereas our evolvability index is inferred from a phylogeny of 25 species. Nevertheless, it is certainly possible that constraints on neutral evolution to some extent translate to constraints on adaptive evolution. Therefore, we take into consideration the dominant factors determining neutral evolutionary rate in order to verify that none of these are in fact responsible for our observed associations between indices of modularity, robustness, and evolvability. We do not find any of them to be confounding. Even though

gene expression level is a dominant factor determining protein evolutionary rate in bacteria (Rocha and Danchin 2004) and yeast (Zhang and He 2005; Drummond et al. 2006), we are not concerned that it is a confounding factor in this study because it has only a negligible role in determining the evolutionary rate of mammalian proteins (Liao et al. 2006; Vinogradov 2010). In fact, because it has only recently been elucidated that the determinants of mammalian protein evolutionary rate differ considerably from those determining the rates in yeast and bacteria, our results are of interest in that they shed preliminary light on how protein structure plays a role in determining the rate of at least adaptive protein evolutionary change in mammals, and they raise the question of whether similar patterns would also be found in bacterial and fungal proteins.

In summary, we conclude that proteins with high rates of adaptive evolution, and thus, high apparent evolvability, have higher helix/strand density and contact density than proteins with lower apparent evolvability and that this pattern is consistent with the idea that modular and/or designable folds—being less structurally constrained—accommodate adaptive changes at a higher rate than proteins with low structural modularity and robustness. Furthermore, we conclude that the effect of structural modularity on protein evolvability is independent of structural robustness and that it is therefore possible that structural modularity drives the relationship between robustness and evolvability observed in proteins.

## Supplementary Material

Supplementary figs. S1–S8 are available at *Genome Biology and Evolution online* (http://gbe.oxfordjournals.org/).

## Acknowledgments

We thank Ben-Yang Liao and Jianzhi Zhang for sharing their data on gene compactness, Trevor Bedford for useful statistical advice, and two anonymous reviewers for their insightful comments and suggestions. This work was supported by the National Institutes for Health (NIH) [5T32GM007499-34]; and the John Templeton Foundation (JTF) [12793]. The views expressed in this paper do not necessarily reflect the views of NIH or JTF.

## Literature Cited

Akiva E, Itzhaki Z, Margalit H. 2008. Built-in loops allow versatility in domain–domain interactions: lessons from self-interacting domains. Proc Natl Acad Sci U S A. 105(36):13292–13297.

Ancel LW, Fontana W. 2000. Plasticity, evolvability, and modularity in RNA. J Exp Zool Part B. 288(3):242–283.

Babajide A, et al. 2001. Exploring protein sequence space using knowledge-based potentials. J Theor Biol. 212:35–46.

Bau D, et al. 2006. Distill: a suite of web servers for the prediction of one-, two- and three-dimensional structural features of proteins. BMC Bioinformatics. 7:402.

Beldade P, Brakefield PM. 2003. Concerted evolution and developmental integration in modular butterfly wing patterns. Evol Dev. 5(2):169–179.

Bhattacharyya RP, Remenyi AR, Yeh BJ, Lim WA. 2006. Domains, motifs, and scaffolds: the role of modular interactions in the evolution and wiring of cell signaling circuits. Annu Rev Biochem. 75:655–680.

Bloom JD, Adami C. 2003. Apparent dependence of protein evolutionary rate on number of interactions is linked to biases in protein–protein interactions data sets. BMC Evol Biol. 3:21.

Bloom JD, Adami C. 2004. Evolutionary rate depends on number of protein–protein interactions independently of gene expression level: response. BMC Evol Biol. 4:14.

Bloom JD, Drummond DA, Arnold FH, Wilke CO. 2006. Structural determinants of the rate of protein evolution in yeast. Mol Biol Evol. 23(9):1751–1761.

Bloom JD, Silberg JJ, Wilke CO, Drummond DA, Adami C. 2005. Thermodynamic prediction of protein neutrality. Proc Natl Acad Sci U S A. 102(3):606–611.

Bogarad LD, Deem MW. 1999. A hierarchical approach to protein molecular evolution. Proc Natl Acad Sci U S A. 96:2591–2595.

Bonner JT. 1988. The evolution of complexity. Princeton (NJ): Princeton University Press.

Brown CJ, et al. 2002. Evolutionary rate heterogeneity in proteins with long disordered regions. J Mol Evol. 55:104–110.

Bustamante C, Townsend JP, Hartl DL. 2000. Solvent accessibility and purifying selection within proteins of Escherichia coli and Salmonella entericca. Mol Biol Evol. 17(2):301–308.

Carbon S, et al. 2009. AmiGO: online access to ontology and annotation data. Bioinformatics. 25(2):288–289.

Chen Y, Dokholyan NV. 2006. The coordinated evolution of yeast proteins is constrained by functional modularity. Trends Genet. 22(8):416–419.

Copley RR, Doerks T, Letunic I, Bork P. 2002. Minireview: protein domain analysis in the era of complete genomes. FEBS Lett. 513:129–134.

Cowperthwaite MC, et al. 2008. The ascent of the abundant: how mutational networks constrain evolution. PLoS Comput Biol. 4(7):e1000110.

Cui Y, Wong WH, Bornberg-Bauer E, Chan HS. 2002. Recombinatoric exploration of novel folded structures: a heteropolymer-based model of protein evolutionary landscapes. Proc Natl Acad Sci U S A. 99(2):809–814.

de Visser AGM, et al. 2003. Perspective: evolution and detection of genetic robustness. Evolution. 57(9):1959–1972.

del Sol A, Arauzo-Bravo MJ, Nussinov R. 2009. Network robustness and modularity of protein structures in the identification of key residues for allosteric communications. J Biol Mol Struct Dyn. 26(6):861–861.

del Sol A, Carbonell P. 2007. The modular organization of domain structures: insights into protein–protein binding. PLoS Comput Biol. 3(12):2446–2455.

Dong J, Horvath S. 2007. Understanding network concepts in modules. BMC Syst Biol. 1:24.

Draghi JA, Parson TL, Wagner GP, Plotkin JB. 2010. Mutational robustness can facilitate adaptation. Nature. 463(7279):353–355.

Drummond DA, Bloom JD, Adami C, Wilke CO, Arnold FH. 2005. Why highly expressed proteins evolve slowly. Proc Natl Acad Sci U S A. 102:14338–14343.

Drummond DA, Raval A, Wilke CO. 2006. A single determinant dominates the rate of yeast protein evolution. Mol Biol Evol. 23:327–337.

Emmert-Streib F, Mushegian A. 2007. A topological algorithm for identification of structural domains of proteins. BMC Bioinformatics. 8:237.

England JL, Shakhnovich EI. 2003. Structural determinants of protein designability. Phys Rev Lett. 90(21):218101.
England JL, Shakhnovich BE, Shakhnovich EI. 2003. Natural selection of more designable folds: a mechanism for thermophilic adaptation. Proc Natl Acad Sci U S A. 100(15):8727–8731.
Ferrada E, Wagner A. 2008. Protein robustness promotes evolutionary innovations on large evolutionary time-scales. Proc R Soc B Biol Sci. 275:1595–1602.
Fontana W. 2002. Modeling 'evo-devo' with RNA. BioEssays. 24:1164–1177.
Force A, et al. 2005. The origin of subfunctions and modular gene regulation. Genetics. 170:433–446.
Franz-Odendaal TA, Hall BK. 2006. Modularity and sense organs in the blind cavefish, Astyanax mexicanus. Evol Dev. 8(1):94–100.
Fraser HB, et al. 2002. Evolutionary rate in the protein interaction network. Science. 296:750–752.
Gardner A, Zuidema W. 2003. Is evolvability involved in the origin of modular variation? Evolution. 57(6):1448–1450.
Garza AS, Ahmad N, Kumar R. 2009. Role of intrinsically disordered protein regions/domains in transcriptional regulation. Life Sci. 84:189–193.
Gerhart J, Kirschner M. 1997. Cells, embryos and evolution. Oxford: Blackwell Science.
Gibson G, Wagner GP. 2000. Canalization in evolutionary genetics: a stabilizing theory? BioEssays. 22(4):372–380.
Govindarajan S, Goldstein RA. 1997. Evolution of Model Proteins on a foldability landscape. Proteins. 29(4):461–466.
Griswold CK. 2006. Pleiotropic mutation, modularity and evolvability. Evol Dev. 8(1):81–93.
Hansen TF. 2002. Is modularity necessary for evolvability? Remarks on the relationship between pleiotropy and evolvability. BioSystems. 69:83–94.
Hartl DL, Taubes CH. 1996. Compensatory nearly neutral mutations: selection without adaptation. J Theor Biol. 182:303–309.
Hartling J, Kim J. 2008. Mutational robustness and geometrical form in protein structures. J Exp Zool Part B. 310B:216–226.
Hartwell LH, Hopfield JJ, Leibler S, Murray AW. 1999. From molecular to modular cell biology. Nature. 402(6761 Suppl):C47–C52.
Herbeck JT, Wall DP. 2005. Converging on a general model of protein evolution. Trends Biotechnol. 23(10):485–487.
Howell DP-G, Samudrala R, Smith JD. 2006. Disguising itself—insights into Plasmodium falciparum binding and immune evasion from the DBL crystal structure. Mol Biochem Parasitol. 148:1–9.
Kabsch W, Sander C. 1983. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers. 22:2577–2637.
Kitano H. 2004. Biological robustness. Nat Rev Genet. 5:826–837.
Krishnan A, Giuliani A, Zbilut JP, Tomita M. 2007. Network scaling invariants help to elucidate basic topological principles of proteins. J Proteome Res. 6(10):3924–3934.
Laborde T, Tomita M, Krishnan A. 2008. GANDivAWeb: a web server for detecting early folding units "foldons") from protein 3D structures. BMC Struct Biol. 8:15.
Li Y, et al. 2007. A diverse family of thermostable cytochrome P450s created by recombination of stabilizing fragments. Nat Biotechnol. 25(9):1051–1056.
Li H, Helling R, Tang C, Wingreen N. 1996. Emergence of preferred structures in a simple model of protein folding. Science. 273:666–669.
Liao B-Y, Scott NM, Zhang J. 2006. Impacts of gene essentiality, expression pattern, and gene compactness on the evolutionary rate of mammalian proteins. Mol Biol Evol. 23(11):2072–2080.
Liao H, Yeh W, Chiang D, Jernigan RL, Lustig B. 2005. Protein sequence entropy is closely related to packing density and hydrophobicity. Protein Eng Des Sel. 18(2):59–64.
Lin Y-S, Hsu W-L, Hwang J-K, Li W-H. 2007. Proportion of solvent-exposed amino acids in a protein and rate of protein evolution. Mol Biol Evol. 24(4):1005–1011.
Lipman DJ, et al. 2002. The relationship of protein conservation and sequence length. BMC Evol Biol. 2:20.
Lipson H, Pollack JB, Suh NP. 2002. On the origin of modular variation. Evolution. 56(8):1549–1556.
Lynch M. 2007. The frailty of adaptive hypotheses for the origins of organismal complexity. Proc Natl Acad Sci U S A. 104:8597–8604.
Marisco A, et al. 2010. Structural fragment clustering reveals novel structural and functional motifs in alpha-helical transmembrane proteins. BMC Bioinformatics. 11:204.
Meiklejohn CD, Hartl DL. 2002. A single mode of canalization. Trends Ecol Evol. 17(10):468–473.
Misevic D, Ofria C, Lenski RE. 2006. Sexual reproduction reshapes the genetic architecture of digital organisms. Proc R Soc B Biol Sci. 273:457–464.
Pereira-Leal JB, Levy ED, Teichmann SA. 2006. The origins and evolution of functional modules: lessons from protein complexes. Philos Trans R Soc Lond B Biol Sci. 361(1467):507–517.
Pigliucci M. 2008. Is evolvability evolvable? Nat Rev Genet. 9:75–82.
Ranwez V, et al. 2007. OrthoMaM: a database of orthologous genomic markers for placental mammal phylogenetics. BMC Evol Biol. 7:241.
Regad L, Martin J, Nuel G, Camproux A-C. 2010. Mining protein loops using a structural alphabet and statistical exceptionality. BMC Bioinformatics. 11:75.
Ridout KE, Dixon CJ, Filatov DA. 2010. Positive selection differs between secondary structure elements in Drosophila. Genome Biol Evol. 2010:166–179.
Rocha EP, Danchin A. 2004. An analysis of determinants of amino acids substitution rates in bacterial proteins. Mol Biol Evol. 21: 108–116.
Schlosser G, Wagner GP. 2004. Modularity in development and evolution. Chicago (IL): University of Chicago Press.
Shakhnovich BE, Deeds E, Delisi C, Shakhnovich E. 2005. Protein structure and evolutionary history determine sequence space topology. Genome Res. 15(3):385–392.
Singh GP, Ganapathi M, Dash D. 2007. Role of intrinsic disorder in transient interactions of hub proteins. Proteins. 66(4):761–765.
Su QJJ, Lu L, Saxonov S, Brutlag DL. 2005. eBLOCKs: enumerating conserved protein blocks to achieve maximal sensitivity and specificity. Nucleic Acids Res. 33:D178–D182.
Sumedha , Martin OC, Wagner A. 2007. New structural variation in evolutionary searches of RNA neutral networks. BioSystems. 90:475–485.
Sreerama N, Venyaminov SYU, Woody RW. 1999. Estimation of the number of a-helical and b-strand segments in proteins using circular dichroism spectroscopy. Protein Sci. 8:360–380.
Taverna DM, Goldstein RA. 2000. The distribution of structures in evolving protein populations. Biopolymers. 53:1–8.
The Gene Ontology Consortium. 2000. Gene Ontology: tool for the unification of biology. Nat Genet. 25:25–29.
Trifonov EN, Frenkel ZM. 2009. Evolution of protein modularity. Curr Opin Struc Biol. 19:335–340.
Vinogradov A. 2010. Systemic factors dominate mammal protein evolution. Proc R Soc B Biol Sci. 277(1686):1403–1408.

Wagner A. 2005. Robustness and evolvability in living systems. Princeton (NJ): Princeton University Press.
Wagner A. 2008. Robustness and evolvability: a paradox resolved. Proc R Soc B Biol Sci. 275:91–100.
Wagner GP. 1996. Homologues, natural kinds, and the evolution of modularity. Am Zool. 36:36–43.
Wagner GP, Altenberg L. 1996. Complex adaptations and the evolution of evolvability. Evolution. 50:967.
Wagner GP, Booth G, Bagheri-Chaichian H. 1997. A population genetic theory of canalization. Evolution. 51:329–347.
Wagner GP, Pavlicev M, Cheverud JM. 2007. The road to modularity. Nat Rev Genet. 8:921–931.
Wilke CO, Bloom JD, Drummond DA, Raval A. 2005. Predicting the tolerance of proteins to random amino acid substitution. Biophys J. 89:3714–3720.
Wright PE, Dyson HJ. 1999. Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. J Mol Biol. 293:321–331.
Wroe R, Chan HS, Bornberg-Bauer E. 2007. A structural model of latent evolutionary potentials underlying neutral networks in proteins. HFSP J. 1(1):79–87.
Xia Y, Levitt M. 2002. Roles of mutation and recombination in the evolution of protein thermodynamics. Proc Natl Acad Sci U S A. 99(16):10382–10387.
Yang AS. 2001. Modularity, evolvability, and adaptive radiations: a comparison of the hemi- and holometabolous insects. Evol Dev. 3(2):59–72.
Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci. 13:555–556.
Yang Z. 2007. PAML 4: a program package for phylogenetic analysis by maximum likelihood. Mol Biol Evol. 24:1586–1591.
Zhang J, He X. 2005. Significant impact of protein dispensability on the instantaneous rate of protein evolution. Mol Biol Evol. 22:1147–1155.

**Associate editor:** George Zhang