doi:10.1038/nature09992

# Non-adaptive origins of interactome complexity

Ariel Fernández[1,2] & Michael Lynch[3]

**The boundaries between prokaryotes, unicellular eukaryotes and multicellular eukaryotes are accompanied by orders-of-magnitude reductions in effective population size, with concurrent amplifications of the effects of random genetic drift and mutation[1]. The resultant decline in the efficiency of selection seems to be sufficient to influence a wide range of attributes at the genomic level in a non-adaptive manner[2]. A key remaining question concerns the extent to which variation in the power of random genetic drift is capable of influencing phylogenetic diversity at the subcellular and cellular levels[2–4]. Should this be the case, population size would have to be considered as a potential determinant of the mechanistic pathways underlying long-term phenotypic evolution. Here we demonstrate a phylogenetically broad inverse relation between the power of drift and the structural integrity of protein subunits. This leads to the hypothesis that the accumulation of mildly deleterious mutations in populations of small size induces secondary selection for protein–protein interactions that stabilize key gene functions. By this means, the complex protein architectures and interactions essential to the genesis of phenotypic diversity may initially emerge by non-adaptive mechanisms.**

Here we examine whether established gene orthologies reveal a role for drift in phylogenetic patterns of protein structural evolution. Although evolutionary change at the structural level is unlikely to destabilize greatly the native fold of an essential protein, as the complete loss of function would generally be unbearable, the drift hypothesis predicts a negative relation between population size ($N$) and the accumulation of mildly deleterious amino-acid substitutions. The following examination of the structures of orthologous proteins from vastly different lineages suggests that the enhanced power of drift in eukaryotes (multicellular species in particular) results in a qualitative reduction in the stability of protein–water interfaces (PWIs) through the partial exposure of paired backbone polar groups (amides and carbonyls) that are otherwise protected in prokaryotes. In effect, the reduced efficiency of selection in small-$N$ species encourages the accumulation of mild structural deficiencies in the form of solvent-accessible backbone hydrogen bonds (SABHBs), which lead to protein structures that are more 'open' and vulnerable to fold-disruptive hydration (Fig. 1a) and create protein–water interfacial tension (PWIT; Supplementary Fig. 1)[5] by hindering the hydrogen-bonding capabilities of nearby water molecules.

We argue that the emergence of unfavourable PWIs promotes the secondary recruitment of novel protein–protein associations that restore structural stability by reducing PWI. Under this hypothesis, complex organisms may frequently develop protein–protein interactions not as immediate vehicles for novel adaptive functions, but as compensatory mechanisms for retaining key gene functions. Once in place, such physical contact between interacting proteins may provide a selective environment for the further emergence of entirely novel protein–protein interactions underlying cellular and organismal complexities. Our suggestion that the hallmark of eukaryotic evolution, the origin of interactome complexity, may have arisen in part as a passive consequence of the enhanced power of drift reduces the need to invoke direct long-term selective advantages of phenotypic complexity[6].

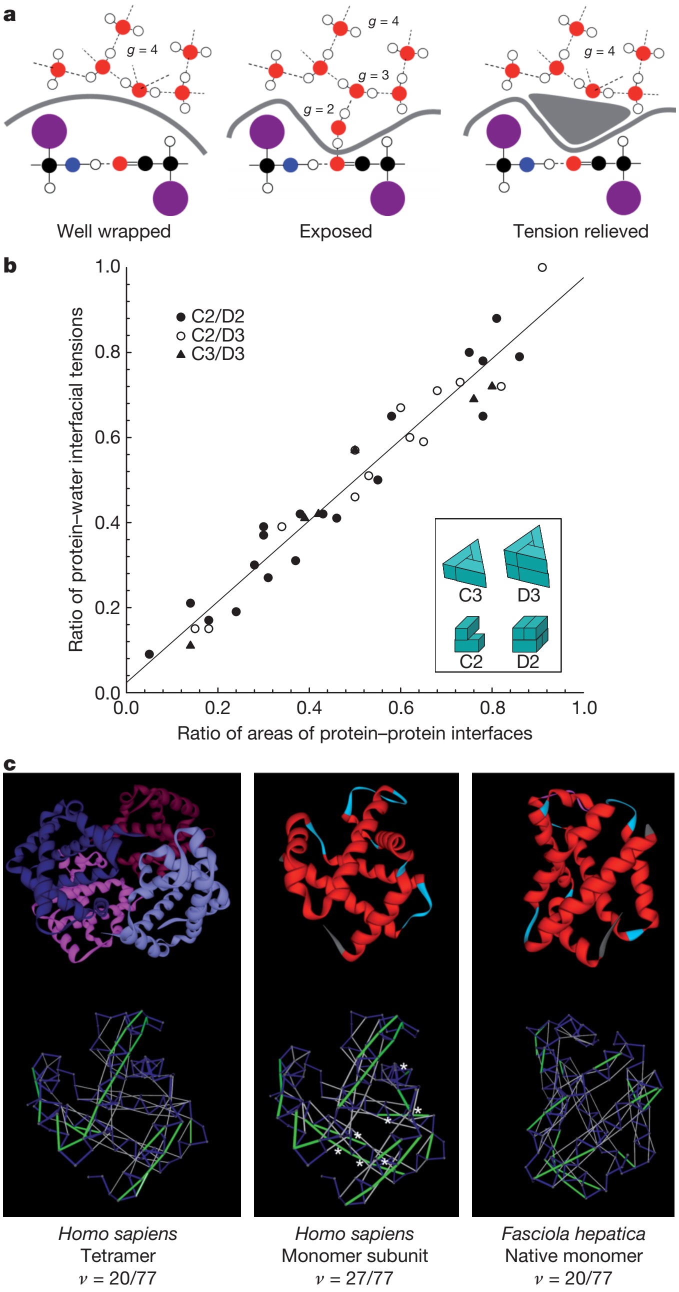
To gain insight into the evolution of interactome complexity, we derived quantitative measures of the PWIT as indicators of potential molecular interactivity. To estimate the PWIT of a protein, we computationally equilibrated the protein structure in surrounding water, using the function $g(\mathbf{r})$ to represent the time-averaged coordination (number of hydrogen bonds) associated with a water molecule at position $\mathbf{r}$ (Fig. 1a), and integrating over the entire protein surface all water molecules within a 10 Å radius (the thickness of four layers of water molecules). Compared with bulk water (where $g = 4$), interfacial water molecules may have reduced hydrogen-bonding opportunities ($g < 4$) and often counterbalance these losses by interacting with polar groups on the protein surface. Thus the PWIT parameter integrates information on unfavourable local decreases in $g$ and favourable polarization contributions from the protein to yield the free-energy cost, $\Delta G_{\mathrm{if}}$, of spanning the protein–water interface (Methods). A high PWIT signals a high propensity for protein–protein associations, which reduce the PWI area.

To validate the use of PWIT as a measure of interactivity, we examined an exhaustive catalogue of contact topologies for protein complexes with one to six subunits, with each topology being evaluated with one or more non-homologous complexes using structures in the Protein Data Bank (PDB) (Supplementary Table 1). For each complex, we computed the total protein–protein interface area after identifying the residues engaged in intermolecular contacts[7]. For each protein subunit, the protein–protein interface is contained within the PWI region that generates tension in the free subunit, and there is a tight correlation between the surface areas for both regions, implying that regions on the protein surface generating PWIT (i.e. those with $g < 4$ for nearby water) actually promote associations (Supplementary Figs 1 and 2a). Next, we verified that protein surface regions generating PWIT coincide with the affinity-contributing regions at protein–protein interfaces. To this end, we tested the value of PWIT as a promoter of protein associations by focusing on the interface for the 1:1 human growth hormone (hGH)-receptor complex[8] (Supplementary Fig. 2b) for which the consequences of amino-acid substitutions have been extensively evaluated. Our analysis reveals a strong correlation between the change in PWIT induced by site-specific mutagenesis of interfacial residues and the association free-energy difference created by the alteration of the hormone–receptor interface (Supplementary Fig. 2c).

Comparison of orthologous proteins engaging in different levels of homo-oligomerization in different species[9] further supports the view that PWIT serves as a measure of the propensity for protein–protein association. The ratio of protein–protein interface areas (lower to higher degrees of complexation; Supplementary Table 2) exhibits a strong positive correlation with the ratio of PWITs for the respective free subunits (Fig. 1b). As complexes with higher degrees of oligomerization arise from lower-order complexes, this implies that the degree of cooperativity among subunits correlates with the PWIT of the basic subunit.

Hydrophobic regions on protein surfaces obviously contribute to PWIT, but analysis of proteins exhibiting association propensity (Supplementary Table 2) shows that the regions generating 73 ± 5%

[1] Department of Computer Science, The University of Chicago, Chicago, Illinois 60637, USA. [2] Department of Bioengineering, Rice University, Houston, Texas 77005, USA. [3] Department of Biology, Indiana University, Bloomington, Indiana 47405, USA.



*Homo sapiens*
Tetramer
$\nu$ = 20/77

*Homo sapiens*
Monomer subunit
$\nu$ = 27/77

*Fasciola hepatica*
Native monomer
$\nu$ = 20/77

**Figure 1 | Structural deficiencies in soluble proteins promote protein associations. a**, Hydration of exposed polar backbone induces interfacial tension by causing water molecules near the defect to relinquish part of their coordination ($g < 4$) relative to the level in surrounding bulk solvent ($g = 4$). White represents hydrogen atoms; red, oxygen; blue, nitrogen; black, carbon; the larger purple circles denote side chains for amino acids. Hydrogen bonds are denoted by dashed lines. Thick grey lines outline the external surface of the overall protein molecule, and the underlying structure represents two amino acids made adjacent by the protein architecture and bound by a hydrogen bond between the backbone amide (blue:white) of one amino acid and carbonyl (red:black) of the other. Water molecules are shown as angular red and white segments, with the coordination number $g$ denoting the number of hydrogen bonds associated with a water molecule ($g = 4$ for bulk water; $g < 4$ for confined interfacial water). In the centre, the structure of the protein causes local exposure and unfavourable hydration of the polar backbone, whereas the absence of such local interactions between water molecules and the well-wrapped proteins on the left and right reduces interfacial tension (interfacial water is bulk-like, retaining the maximum coordination $g = 4$). **b**, Comparison of orthologous proteins with different levels of homo-oligomerization reveals that the PWIT is an indicator of the propensity for cooperative improvement/refinement of protein function through complexation. The ratio of protein–protein interfaces (small to large) was determined for pairs of orthologous proteins with different levels of oligomerization in different species (Supplementary Table 2) and plotted against the ratio of PWITs for the respective free subunits. The tight correlation ($r^2 = 0.94$) reveals that interspecific differences in PWIT accompany differences in levels of oligomerization, thus providing a measure of potential allosteric or cooperative improvement of basic protein function. Complexes with cyclic rotational symmetry (C2, C3, …) can further oligomerize into complexes with dihedral (D2, D3, …) symmetry, as shown in the idealized diagrams in the lower right. For example, C2 complexes can dimerize into D2 complexes, trimerize into D3 complexes, etc., whereas a D3 complex can also be obtained by dimerization of a C3 complex. For the protein–protein interface and PWIT ratios examined, the interface for the subunit in the complex with lower-order symmetry is compared with that in the complex with higher-order symmetry, yielding analyses based on protein pairs contrasted within three groupings: C2 versus D2, C2 versus D3, and C3 versus D3. **c**, The SABHB patterns from two haemoglobins with different oligomerization levels in their native states are compared. In the bottom panels, the protein backbone is represented by virtual bonds in blue joining α-carbons, with well-protected BHBs shown as light grey and SABHBs as green lines joining the α-carbons of the paired residues. The ribbon representations of the human complex and dissociated subunit (chain A in PDB.2DN2, left and centre, respectively) are included as aids to the eye, representing the structuring of the backbone in each subunit. The free subunit isolated from the tetramer in *H. sapiens* (PDB.2DN2, chain A, centre) has seven excess SABHBs (denoted by stars) when compared with the subunit within the tetrameric complex, where they are well-protected intermolecularly, alleviating interfacial tension. As a consequence of this better wrapping, the overall extent of structural deficiency ($\nu$ value) for the subunit within the human complex is identical to that of the natively monomeric haemoglobin from the trematode *F. hepatica* (PDB.2VYW). This raises the possibility that the accumulation of structural deficiencies in the mammalian haemoglobin subunit promoted the emergence of an oligomeric association as a means of reducing excess interfacial tension. The structural displays were obtained by uploading the PDB text files into the program YAPview, a displayer of local backbone desolvation of soluble proteins that can be downloaded from the link 'Dehydron Calculator' at http://www.owlnet.rice.edu/~arifer/.

of the PWIT (Supplementary Information and Supplementary Fig. 1) arise from SABHBs. The resultant hydration of backbone polar groups (amides and carbonyls) causes a loss of coordination for local water molecules, which increases surface tension and creates an unstable PWI, as the cavities cannot accommodate a bulk-like water molecule[10]. As an example of how such a structural deficiency can be alleviated through a protein association, an isolated α-subunit of the human haemoglobin tetramer has seven SABHBs that become protected within the tetrameric complex, such that the ratio ($\nu$) of SABHBs to total BHBs in the complex-associated subunit is the same as that in the natively monomeric unit for haemoglobin from the trematode *Fasciola hepatica* (Fig. 1c).

To evaluate whether the accumulation of structural deficiencies of proteins is generally encouraged by random genetic drift, and in turn enhances the propensity for establishing protein complexes, we examined a set of 106 orthologous water-soluble proteins (sequence identity greater than 30%)[11,12] with PDB-reported structures for at least two species. We considered 36 species with vastly different population sizes[1,2], each containing proteins in at least 90 of the 106 orthologous groups (Supplementary Tables 3–5). Template-based three-dimensional structures for orthologues lacking PDB-reported structures were constructed by homology threading[13,14], and evaluated, ranked and selected according to the energetic proximity between template and model[15]. The accuracy of this homology-based prediction of PWIT was determined with a test set of proteins with PDB-reported structures from two species, subjecting one member of each orthologous pair to homology threading through the other. Comparison of the indirect and direct estimates of PWIT demonstrates that when sequence identities are greater than 35%, the predicted PWIT diverges less than 10% from the more direct estimate for the same protein (Supplementary Fig. 3).

For each protein structure, $g(\mathbf{r})$ was obtained as described in Methods, and the relative propensities for protein association across orthologues were then determined by assessing differences in the free-energy cost

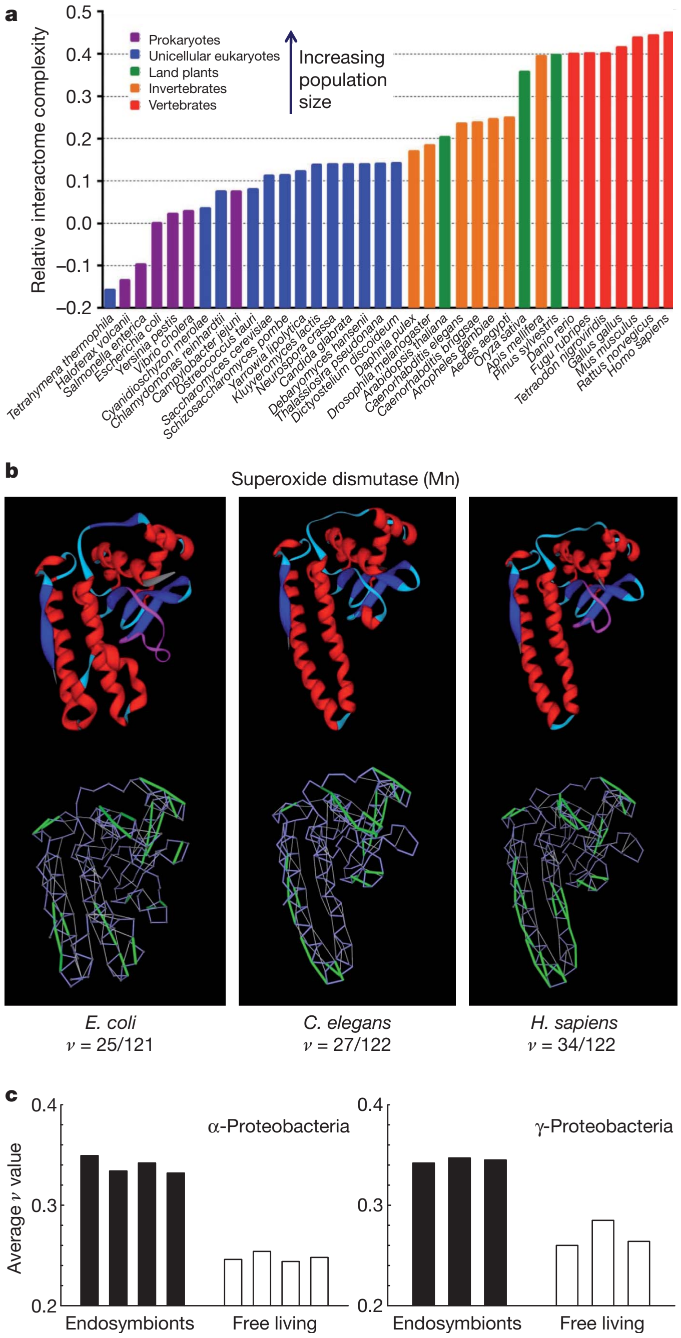


**Figure 2 | Structural degradation enhances PWIT and promotes protein interactivity in species with small population sizes.** **a**, Potential for interactome complexity of 36 species with diverse population sizes (Supplementary Table 2), relative to *E. coli*. To highlight the relative power of random genetic drift, bars are colour-coded to reflect groupings of species in broad population-size categories. **b**, Overall structural deficiency of orthologues of the enzyme superoxide dismutase (Mn), revealing a progressive accumulation of SABHBs in the orthologues of the bacterium *E. coli*, the nematode *Caenorhabditis elegans* and *H. sapiens*. The upper ribbon representations illustrate the structural conservation across orthologues (respective PDB accession numbers 3ot7, 3dc6, 2adq). The conventional colour coding is red, blue, magenta and light blue for helix, β-strand, loop and turn, respectively. **c**, Average structural deficiency ($\nu$ value) of protein orthologues for intracellular and free-living bacterial species. Species identities, progressing from left to right are as follows: α-Proteobacteria—*Rickettsia typhi*, *Orientia tsutsugamushi*, *Anaplasma centrale* str. Israel, *Wolbachia* sp. wRi, *Rhodospirillum centenum* SW, *Magnetospirillum magneticum*, *Silicibacter TM1040*, *Erythrobacter litoralis*; γ-Proteobacteria—*Buchnera aphidicola*, *Wigglesworthia brevipalpis*, Candidatus *Blochmannia pennsylvanicus*, *Marinomonas* MWYL1, *E. coli*, *Pseudomonas aeruginosa*. Only proteins with orthologues across the full set of species within each group were considered for analysis (Supplementary Tables 6 and 7).

$\Delta G_{\text{if}}$ among species. We estimated the relative complexation propensity $M_{j,n}$ of a protein in orthologue group $j$ (1, …, 106) from species $n$ (1, …, 36) by adopting *Escherichia coli* as a reference species ($n = 1$):

$$M_{j,n} = [(\Delta G_{\text{if}})_{j,n} - (\Delta G_{\text{if}})_{j,1}]/(\Delta G_{\text{if}})_{j,1} \qquad (1)$$

With this index, $M_{j,1} = 0$ for all proteins in *E. coli*, and taxa with less well-wrapped proteins (and hence greater propensity for complexation), have positive values.

The mean value of species-specific estimates of $M_{j,n}$ over all proteins evaluated is negatively correlated with the approximate effective population sizes of species (Fig. 2a), given that the average ranking of the latter is prokaryotes > unicellular eukaryotes > invertebrates > vertebrates and land plants[1,2]. A specific example of a trend towards increasing structural openness with reduced population sizes is illustrated in Fig. 2b, where the SABHB patterns and $\nu$ values for orthologues of the enzyme superoxide dismutase are compared across three species.

The results from Fig. 2a and an additional analysis (Supplementary Fig. 4) support the hypothesis that large organisms with small population sizes experience a significant enough increase in the power of random genetic drift to magnify the accumulation of mild structural deficiencies in the form of SABHBs, resulting on average in proteins with a more solvent-exposed or 'open' structure. By contrast, mutations to SABHBs are more frequently excluded by selection in species with larger population sizes (for example, prokaryotes). Thus, because SABHBs are the main determinants of interfacial tension (Supplementary Fig. 1), the proteins of large organisms have a greater inherent tendency to form novel protein–protein associations (Fig. 1a). This suggests that increases in protein-network complexity in multicellular species may in part owe their origins to modifications to the intracellular selective environment induced by non-adaptive structural degradation of individual proteins.

One concern with the preceding interpretation is the order of events: does an initial degradation of architectural integrity of individual proteins in response to random genetic drift induce secondary selection for the recruitment of interacting partners, or does the emergence of cellular complexity (and increased protein interactivity) precede secondary changes in protein sequence to accommodate such interactions? One way to evaluate this matter is to compare proteins from related species that have experienced relatively recent divergences in effective population sizes but no major modifications in intracellular complexity or emergence of multicellularity.

To achieve this task, we compared orthologous genes from endosymbiotic/intracellular bacteria and their free-living relatives, as the former are thought to have experienced substantial reductions in effective population sizes[16]. Previous suggestions that intracellular bacteria experience elevated levels of random genetic drift have been based on ratios of substitution rates at silent and replacement sites, which can be biased indicators of the efficiency of selection if there is selection on silent sites. Although the lack of protein structural information for endosymbiotic species requires a sequence-based identification of SABHBs derived from reliable scores of native disorder propensity (Methods), the resultant analyses are broadly consistent with the hypothesis that an increase in the power of drift in microbes encourages the accumulation of structural defects in protein architecture (Fig. 2c). Free-living species, with larger effective population sizes, have consistently smaller $\nu$ values for orthologous genes in both α- and γ-Proteobacteria. (Application of the same sort of analysis of disorder propensity across a set of 105 species and 541 proteins corroborates this result (Supplementary Figs 5–7).)

Taken together, our analyses support the hypothesis that the range of population sizes experienced by natural populations is sufficient to induce significantly different patterns of evolution at the level of protein architecture. The resultant changes in the intracellular environment in small-$N$ species provides an opportunity for the recruitment of stabilizing protein–protein interactions, yielding a plausible mechanism for the emergence of molecular complexities before their exploitation in phenotypic divergence[9,17]. This hypothesis does not deny a potentially significant role for natural selection in using such novelties subsequent to their establishment, nor does it deny the fact that intramolecular compensatory mutations can alleviate some structural defects associated with SABHBs. However, our results do raise questions about the necessity of invoking an intrinsic advantage to organismal complexity, and provide a strong rationale for expanding comparative studies in molecular evolution beyond linear sequence analysis to evaluations of molecular structure.

## METHODS SUMMARY

We determined the propensity of proteins to be engaged in associations that reduce the PWI by computing the PWIT. This thermodynamic parameter gives $\Delta G_{\mathrm{if}}$, the free-energy cost of spanning the PWI. The PWIT is computed as

$$\Delta G_{\mathrm{if}} = \tfrac{1}{2}\int\{a|\nabla g|^2 - |\mathbf{P}[g(\mathbf{r})]|^2\}\mathrm{d}\mathbf{r}, \qquad (2)$$

where the term $\tfrac{1}{2}a|\nabla g|^2$, with a = 9.02 mJ m$^{-1}$ at $T$ = 298 K (Methods), accounts for tension-generating reductions in water coordination, and the polarization $\mathbf{P}[g(\mathbf{r})]$ accounts for dipole–electrostatic field interactions (Methods). For a given protein structure or template-based structural model, the field $g = g(\mathbf{r})$ used in the numerical integration of equation (2) was determined by equilibrating the water-embedded structure within an isothermal–isobaric (NPT) ensemble (with fixed parameters $N$ = number of particles, $P$ = pressure and $T$ = temperature; Methods)[10,18,19]. From structural coordinates, we determined the structural deficiencies (SABHBs)[20] that generate 73±5% of the PWIT (Supplementary Information). We examined 106 groups of orthologous proteins identified using OrthoMCL[11,12] for which there are PDB representatives from at least two species (usually *E. coli* and *Homo sapiens*, Supplementary Tables 3–5). We considered 36 representative species, each containing proteins in at least 90 of the 106 orthologue groups. Template-based three-dimensional structures for orthologues lacking a PDB-reported structure[14] were constructed using MODELLER[13], with side chains directly positioned with SCWRL[21]. The template and resulting model were evaluated, ranked and finally selected using ProSA[15]. The accuracy of homology models is shown in Supplementary Fig. 3. In cases where orthologous structural templates were unavailable, like the comparison of endosymbionts with free-living species, a sequence-based inference of SABHBs was performed based on an established anti-correlation between backbone protection and disorder propensity (Supplementary Fig. 5)[22]. The cross validation of homology- and disorder-based estimations of $v$ values is given in Supplementary Fig. 6.

**Full Methods** and any associated references are available in the online version of the paper at www.nature.com/nature.

**Received 24 August 2010; accepted 9 March 2011.**

**Published online 18 May 2011.**

1. Lynch, M. & Conery, J. S. The origins of genome complexity. *Science* **302,** 1401–1404 (2003).
2. Lynch, M. *The Origins of Genome Architecture* (Sinauer, 2007).
3. Stoltzfus, A. On the possibility of constructive neutral evolution. *J. Mol. Evol.* **49,** 169–181 (1999).
4. Gray, M. W., Lukes, J., Archibald, J. M., Keeling, P. J. & Doolittle, W. F. Cell biology. Irremediable complexity? *Science* **330,** 920–921 (2001).
5. Rowlinson, J. S. & Widom, B. *Molecular Theory of Capillarity* (Oxford Univ. Press, 1982).
6. Lynch, M. The frailty of adaptive hypotheses for the origins of organismal complexity. *Proc. Natl Acad. Sci. USA* **104** (Suppl.), 8597–8604 (2007).
7. Levy, E. D., Pereira-Leal, J. B., Chothia, C. & Teichmann, S. A. 3D Complex: a structural classification of protein complexes. *PLOS Comput. Biol.* **2,** e155 (2006).
8. Clackson, T., Ultsch, M. H., Wells, J. A. & de Vos, A. M. Structural and functional analysis of the 1:1 growth hormone:receptor complex reveals the molecular basis for receptor affinity. *J. Mol. Biol.* **277,** 1111–1128 (1998).
9. Levy, E. D., Boeri Erba, E., Robinson, C. V. & Teichmann, S. A. Assembly reflects evolution of protein complexes. *Nature* **453,** 1262–1265 (2008).
10. Fenimore, P. W., Frauenfelder, H., McCammon, B. H. & Young, R. D. Bulk solvent and hydration-shell fluctuations, similar to α- and β-fluctuations in glasses, control protein motions and functions. *Proc. Natl Acad. Sci. USA* **101,** 14408–14413 (2004).
11. Ostlund, G. *et al.* InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. *Nucleic Acids Res.* **38,** D196–D203 (2010).
12. Gabaldon, T. *et al.* Joining forces in the quest for orthologs. *Genome Biol.* **10,** 403 (2009).
13. Sali, A. & Blundell, T. L. Comparative protein modeling by satisfaction of spatial restraints. *J. Mol. Biol.* **234,** 779–815 (1993).
14. Zhou, H. & Skolnick, J. Improving threading algorithms for remote homology modeling by combining fragment and template comparisons. *Proteins* **78,** 2041–2048 (2010).
15. Wiederstein, M. & Sippl, M. J. ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins. *Nucleic Acids Res.* **35,** W407–W410 (2007).
16. Moran, N. A. Accelerated evolution and Muller's ratchet in endosymbiotic bacteria. *Proc. Natl Acad. Sci. USA* **93,** 2873–2878 (1996).
17. Kuriyan, J. & Eisenberg, D. The origin of protein interactions and allostery in colocalization. *Nature* **450,** 983–990 (2007).
18. Rizzo, R. C. & Jorgensen, W. L. OPLS All-atom model for amines: resolution of the amine hydration problem. *J. Am. Chem. Soc.* **121,** 4827–4836 (1999).
19. Jorgensen, W. L., Chandrasekhar, J., Madura, J., Impey, R. W. & Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **79,** 926–935 (1983).
20. Fernández, A. & Berry, R. S. Golden rule for buttressing vulnerable soluble proteins. *J. Proteome Res.* **9,** 2643–2648 (2010).
21. Canutescu, A. A., Shelenkov, A. & Dunbrack, R. L. A graph-theory algorithm for rapid protein side-chain prediction. *Protein Sci.* **12,** 2001–2014 (2003).
22. Pietrosemoli, N., Crespo, A. & Fernández, A. Dehydration propensity of order-disorder intermediate regions in soluble proteins. *J. Proteome Res.* **6,** 3519–3526 (2007).

**Supplementary Information** is linked to the online version of the paper at www.nature.com/nature.

**Acknowledgements** A.F. was supported by National Institutes of Health grant R01GM072614, and by the Institute of Biophysical Dynamics and the Department of Computer Science at The University of Chicago. M.L. was supported by National Institutes of Health grant R01GM036827 and National Science Foundation grant EF-0827411.

**Author Contributions** A.F. and M.L. conceived the project and wrote the paper. A.F. collected the orthologue groups across 36 species with sufficient structural representation, performed the structural analysis and determined the interaction propensities across orthologues.

**Author Information** Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to M.L. (milynch@indiana.edu) or A.F. (ariel@uchicago.edu).

## METHODS

**Computation of PWIT.** The parameter $a$ in equation (2) is obtained from the interfacial tension of a large non-polar sphere with radius $\theta$ in the limit $\theta/1\,\mathrm{nm} \rightarrow \infty$. Thus we get $a = 9.02\,\mathrm{mJ\,m^{-1}} = \mathrm{limit}_{\theta/1\,\mathrm{nm}\rightarrow\infty}[\gamma(4\pi\theta^2)/\int \tfrac{1}{2}|\nabla g|^2 d\mathbf{r}]$, where $\gamma = 72\,\mathrm{mJ\,m^{-2}}$ is the bulk surface tension of water at 298 K, and $\int|\nabla g|^2 d\mathbf{r} = O(\theta^2)$ since $\nabla g \neq 0$ only in the vicinity of the interface. To determine the $g$-dependence of polarization $\mathbf{P} = \mathbf{P}(\mathbf{r})$, we adopt the Fourier-conjugate frequency space ($\boldsymbol{\omega}$ space) and represent the dipole correlation kernel $K_\mathrm{p}(\boldsymbol{\omega})$ and the electrostatic field $\mathbf{E} = \mathbf{E}(\mathbf{r})$ in this space. In contrast with other treatments[23], we note that $\mathbf{P}$ and $\mathbf{E}$ are indeed proportional but the proportionality constant is $\boldsymbol{\omega}$ dependent[24]. Thus, in $\boldsymbol{\omega}$ space, we get

$$F(\mathbf{P})(\boldsymbol{\omega}) = K_\mathrm{p}(\boldsymbol{\omega})F(\mathbf{E})(\boldsymbol{\omega}), \tag{3}$$

where $F$ denotes three-dimensional Fourier transform $F(\mathbf{f})(\boldsymbol{\omega}) = (2\pi)^{-3/2}\int e^{i\boldsymbol{\omega}.\mathbf{r}}\,\mathbf{f}(\mathbf{r})d\mathbf{r}$, and the kernel $K_\mathrm{p}(\boldsymbol{\omega})$ is the Lorentzian $K_\mathrm{p}(\boldsymbol{\omega}) = (\varepsilon_\mathrm{b} - \varepsilon_\mathrm{o})/(1 + (\tau(\mathbf{r})c)^2|\boldsymbol{\omega}|^2)$, with $\tau(\mathbf{r})\mathrm{c}$ = position-dependent dielectric relaxation scale $\approx$ 3 cm for $\tau = \tau_\mathrm{b} \approx 100$ ps ($c$ = speed of light), $\varepsilon_\mathrm{b}$ = bulk permittivity and $\varepsilon_\mathrm{o}$ = vacuum permittivity. Because $\mathbf{P}(\mathbf{r})$ satisfies the Debye relation $\nabla.(\varepsilon_\mathrm{o}\mathbf{E} + \mathbf{P})(\mathbf{r}) = \rho(\mathbf{r})$, where $\rho(\mathbf{r})$ = charge density, equation (4) yields the following equation in $\mathbf{r}$ space[25]:

$$\nabla.[\int F^{-1}(\mathrm{K})(\mathbf{r} - \mathbf{r}')\mathbf{E}(\mathbf{r}')d\mathbf{r}')] = \rho(\mathbf{r}), \tag{4}$$

with $K(\boldsymbol{\omega}) = \varepsilon_\mathrm{o} + K_\mathrm{p}(\boldsymbol{\omega})$. The convolution $\int F^{-1}(\mathrm{K})(\mathbf{r} - \mathbf{r}')\mathbf{E}(\mathbf{r}')d\mathbf{r}'$ captures the correlation of the dipoles with the electrostatic field. Note that equation (4) is not the Poisson–Boltzmann equation, which requires a proportionality between the fields $\mathbf{E}$ and $\mathbf{P}$ under the *ad hoc* assumption $K(\boldsymbol{\omega}) \equiv$ constant.

Upon water confinement, the dielectric relaxation undergoes a frequency redshift arising from the reduction in hydrogen-bond partnerships that translates to a reduction in dipole orientation possibilities. Thus, at position $\mathbf{r}$, the relaxation time is $\tau = \tau_\mathrm{b}\exp(B(g(\mathbf{r}))/k_BT)$, where the kinetic barrier $B(g(\mathbf{r})) = -k_BT\ln(g(\mathbf{r})/4)$ yields $\tau(\mathbf{r}) = \tau_\mathrm{b}(g(\mathbf{r})/4)^{-1}$. Thus, for charge distribution,

$$\rho(\mathbf{r}) = \Sigma_{m\in L}4\pi q_m\delta(\mathbf{r} - \mathbf{r}_m), \tag{5}$$

with $L$ = set of charges on the protein surface labelled by index $m$, the $g$-dependent polarization is obtained from equation (4) (Supplementary Information):

$$\mathbf{P}(\mathbf{r}) = \int F^{-1}(K_\mathrm{p})(\mathbf{r} - \mathbf{r}')\mathbf{E}(\mathbf{r}')d\mathbf{r}'$$

$$= (2\pi)^{-3}\boldsymbol{\Sigma}_{m\in L}\int d\mathbf{r}'F^{-1}(K_\mathrm{p})(\mathbf{r} - \mathbf{r}')\nabla_{\mathbf{r}'}\int d\boldsymbol{\omega} e^{-i\boldsymbol{\omega}.(\mathbf{r}' - \mathbf{r}_m)}4\pi q_m/[|\omega|^2K(\boldsymbol{\omega})]. \tag{6}$$

**Spatially dependent coordination $g = g(\mathbf{r})$.** The time-averaged scalar field $g = g(\mathbf{r})$ was obtained from classical trajectories generated by molecular dynamics. The computations started with the PDB structure of a free (uncomplexed) protein molecule embedded in a pre-equilibrated cell of explicitly represented water molecules and counterions[18,19]. The molecular-dynamics trajectories were generated by adopting an integration time step of 2 fs in an NPT ensemble with box size $10^3$ nm$^3$ and periodic boundary conditions[26]. The box size was calibrated so that the solvation shell extended at least 10 Å from the protein surface at all times. The long-range electrostatics were treated using the particle mesh Ewald summation method[27]. A Nosé–Hoover thermostat[28] was used to maintain the temperature at 300 K, and a Tip3P water model with the optimized potential for liquid simulations (OPLS) force field was adopted[18,19]. A barostat scheme was maintained through a dedicated routine with the pressure held constant at 1 atm. using a weak-coupling algorithm[29]. After equilibration for 300 ns, $g$ values averaged over a time span of 100 ns were determined for each point in space.

**PWIT as promoter of protein–protein associations.** The PWIT computed using equations (2) and (6) is generated by interfacial hotspots of red-shifted dielectric relaxation ($g(\mathbf{r}) < 4$, $\tau(\mathbf{r}) > \tau_\mathrm{b}$). The most common spots involve hindered polar hydration generated by SABHBs (Fig. 1a). Taken collectively, the SABHBs contribute 73 ± 5% to the interfacial tension (Supplementary Information). The results are validated by showing that the inferred patches of interfacial tension promote protein associations, a conclusion supported by the tight correlation ($r^2 = 0.83$) between the total area of surface patches begetting PWIT (increasing the value of the integral in equation (2)) in free complex subunits, and the total protein–protein interfacial area of protein complexes (Supplementary Fig. 2a). The relevance of PWIT as a molecular determinant of protein–protein interactions is further validated by showing that inferred tension patches actually coincide with hotspots at complex interfaces experimentally identified by mutational scanning (Supplementary Fig. 2b, c).

**Identification of SABHBs in soluble proteins.** The extent of protection of a backbone hydrogen bond, $\zeta$, was computed directly from PDB structural coordinates by determining the number of side-chain non-polar groups contained within a desolvation domain around the bond[20,22]. This domain was defined as two intersecting spheres of fixed radius (approximate thickness of three water layers) centred at the α-carbons of the residues paired by the hydrogen bond. In structures of soluble proteins, backbone hydrogen bonds are protected on average by $\zeta = 26.6 \pm 7.5$ non-polar groups for a desolvation sphere of radius 6 Å. SABHBs lie in the tails of the distribution: that is, their microenvironment contains 19 or fewer non-polar groups ($\zeta \leq 19$), so their $\zeta$ value is below the mean minus one standard deviation.

**Sequence-based identification of SABHBs.** SABHBs represent structural vulnerabilities that have been characterized as belonging to a twilight zone between order and native disorder. This characterization is justified by a strong correlation between intramolecular hydrogen-bond protection, $\zeta$, and propensity for structural disorder ($f_\mathrm{d}$) (Supplementary Fig. 5). The correlation reveals that the inability to exclude water intramolecularly from pre-formed hydrogen bonds is causative of the loss of structural integrity. The disorder propensity is accurately quantified by a sequence-based score generated by the program PONDR-VLXT[30], a predictor of native disorder that takes into account residue attributes such as hydrophilicity, aromaticity and their distribution within the window interrogated. The disorder score ($0 \leq f_\mathrm{d} \leq 1$) is assigned to each residue within a sliding window, representing the predicted propensity of the residue to be in a disordered region ($f_\mathrm{d} = 1$, certainty of disorder; $f_\mathrm{d} = 0$, certainty of order). Only 6% of 1,100 non-homologous PDB proteins gave false-positive predictions of disorder in sequence windows of 40 amino acids[22,30]. The strong correlation (Supplementary Fig. 5) between the disorder score of a residue and extent of protection of the hydrogen bond engaging the residue (if any) provides a sequence-based method of inference of SABHBs and supports the picture that such bonds belong to an order–disorder twilight zone[22]. Thus SABHBs can be safely inferred in regions where the disorder score lies in the range $0.35 \leq f_\mathrm{d} < 0.95$, which corresponds to a marginal BHB protection with $7 \leq \zeta \leq 19$ (Supplementary Fig. 5).

**Evaluation of homology models.** The homology models based on template PDB structures from orthologous proteins were evaluated, ranked and ultimately selected using ProSA[15], based on the minimization of $(Z_\mathrm{mod} - Z_\mathrm{temp})/Z_\mathrm{temp}$, where $Z_\mathrm{mod}$ and $Z_\mathrm{temp}$ are the $Z$ scores of model and template. The $Z$ score of a structure or template-based model is the energetic gap between the structure and an average over an ensemble of random conformations for the protein chain[15].

23. Schutz, C. N. & Warshel, A. What are the dielectric constants of proteins and how to validate electrostatic models? *Proteins Struct. Funct. Genet.* **44,** 400–417 (2001).
24. Scott, R., Boland, M., Rogale, K. & Fernández, A. Continuum equations for dielectric response to macromolecular assemblies at the nanoscale. *J. Phys. A* **37,** 9791–9803 (2004).
25. Fernández, A., Sosnick, T. R. & Colubri, A. Dynamics of hydrogen-bond desolvation in folding proteins. *J. Mol. Biol.* **321,** 659–675 (2002).
26. Lindahl, E., Hess, B. & Van der Spoel, D. GROMACS 3.0: a package for molecular simulation and trajectory analysis. *J. Mol. Model.* **7,** 302–317 (2001).
27. Darden, T., York, D. & Pedersen, L. Particle mesh Ewald: an *N* log(*N*) method for Ewald sums in large systems. *J. Chem. Phys.* **98,** 10089–10092 (1993).
28. Hoover, W. G. Canonical dynamics: equilibrium phase-space distributions. *Phys. Rev. A* **31,** 1695–1697 (1985).
29. Berendsen, H. J., Postma, J. P., van Gunsteren, W. F., DiNola, A. & Haak, J. R. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* **81,** 3684–3690 (1984).
30. Li, X., Romero, P., Rani, M., Dunker, A. K. & Obradovic, Z. Predicting protein disorder for N-, C-, and internal regions. *Genome Informat.* **10,** 30–40 (1999).