BIOINFORMATICS ORIGINAL PAPER Vol. 26 no. 18 2010, pages 2266–2272
doi:10.1093/bioinformatics/btq412

Structural bioinformatics

Advance Access publication August 2, 2010

Exposing the co-adaptive potential of protein–protein interfaces through computational sequence design

Menachem Fromer[1] and Michal Linial[2,*]

[1]School of Computer Science and Engineering and [2]Department of Biological Chemistry, Institute of Life Sciences, Sudarsky Center for Computational Biology, The Hebrew University of Jerusalem, Jerusalem, Israel

Associate Editor: Anna Tramontano

ABSTRACT

Motivation: In nature, protein–protein interactions are constantly evolving under various selective pressures. Nonetheless, it is expected that crucial interactions are maintained through compensatory mutations between interacting proteins. Thus, many studies have used evolutionary sequence data to extract such occurrences of correlated mutation. However, this research is confounded by other evolutionary pressures that contribute to sequence covariance, such as common ancestry.

Results: Here, we focus exclusively on the compensatory mutations deriving from physical protein interactions, by performing large-scale computational mutagenesis experiments for >260 protein–protein interfaces. We investigate the potential for co-adaptability present in protein pairs that are always found together in nature (obligate) and those that are occasionally in complex (transient). By modeling each complex both in bound and unbound forms, we find that naturally transient complexes possess greater relative capacity for correlated mutation than obligate complexes, even when differences in interface size are taken into account.

Contact: michall@cc.huji.ac.il

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on April 27, 2010; revised on June 15, 2010; accepted on July 06, 2010

1 INTRODUCTION

Regions that are important for the function and structure of proteins tend to be highly conserved. Thus, many methods have been developed to measure functional and structural constraints in proteins. Identifying correlated mutations (CMs) is one of the most direct measures for revealing the evolutionary constraints that have shaped protein structures and their interaction specificity (Deeds *et al*., 2007). The idea is simple: for mutations to survive purifying selection after one protein has mutated, the fitness of its partner proteins must be rescued through compensatory mutations. Thus, a pair of positions is considered to have a CM if the amino acid identities at the position in the first protein are correlated with the amino acid identities in the partner protein (Thomas *et al*., 2009). Several computational approaches have been developed to predict coevolved residues (Fariselli *et al*., 2001; Pazos *et al*., 1997). CM pairs have been observed both at the intra-molecular (Shackelford and Karplus, 2007; Thomas *et al*., 2009) and inter-molecular (Thomas *et al*., 2009; Weigt *et al*., 2009) levels. However, the CM signal is often too weak to be detected, as it is masked by the large number of non-coevolving residues (Dunn *et al*., 2008). Thus, the study of CM in proteins is often based on some assumption regarding the mechanisms that have led to the CM, such as the rate of mutations, evolutionary time scale, distances between the residues and more (Capra and Singh, 2008), which has often led to a case-by-case understanding (Jothi *et al*., 2006).

To overcome the low signal of CM, some studies focused on very large superfamilies, such as G-protein coupled receptors (Oliveira *et al*., 2002) or hemoglobin (Pazos *et al*., 1997). However, Halperin *et al.* (2006) concluded that current methodologies for detecting CM are not suitable for large-scale inter-molecular contact prediction. One explanation for this poor performance is that it is difficult to separate the effects of physical interactions within protein complexes (co-adaptation) from other forces that can also create global patterns of co-evolution (Chi *et al*., 2008; Hakes *et al*., 2007; Kann *et al*., 2009; Pazos and Valencia, 2008). Nonetheless, some recent research does improve these results through various normalization and filtration schemes (Dunn *et al*., 2008; Kundrotas and Alexov, 2006; Lee and Kim, 2009; Yeang and Haussler, 2007). In addition, by analyzing lattice model proteins it was found that CM patterns are often consistent with the requirement for thermal stability (Berezovsky *et al*., 2007).

On the other hand, the rapid growth in the number of available 3D protein complexes provides a unique opportunity to extract statistical trends for the interfaces of protein complexes (Ansari and Helms, 2005; Ofran and Rost, 2003; Ponstingl *et al*., 2000). Such statistics have been used, e.g. for improving predictive docking (Madaoui and Guerois, 2008; Smith and Sternberg, 2002). In the work of (Mintseris and Weng, 2005), it was shown that protein interfaces are indeed under selection that can be traced by a CM approach. In that study, a large set of heteromeric complexes was carefully compiled and manually partitioned into transient and obligate interactions. The authors found that the interfaces of transient complexes have very little signal of CM, whereas obligate complexes show strong trends of compensatory mutations.

In this study, our goal was to leverage the power of large structural datasets and recent advances in efficient modeling of protein structures to focus exclusively on the co-adaptation resulting

*To whom correspondence should be addressed.

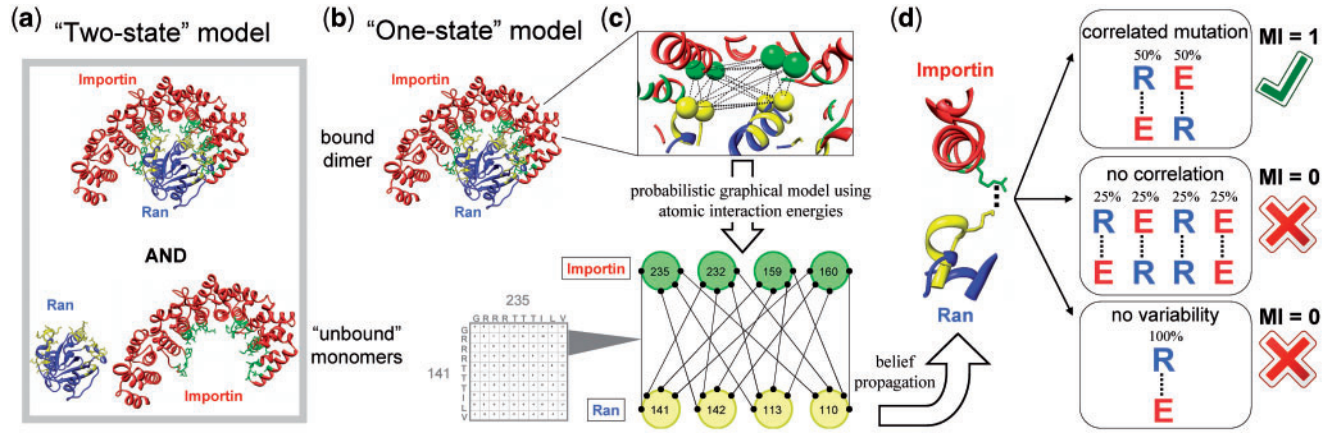

Fig. 1. Structural modeling of correlated mutations in transient and obligate protein–protein complexes. Modeling of the (**a**) two-state and (**b**) one-state scenarios using multi-state computational protein design is illustrated for the Ran-Importin complex (PDB code 1IBR). Interface positions are colored in yellow and green for Ran and Importin, respectively. (**c**) Each structure is modeled as a graph, where nodes correspond to residues and the edges between them contain their atomic-level energetic interactions summed over all atoms; for simplicity, only inter-molecular edges are depicted. The belief propagation algorithm is used to derive globally consistent amino acid probabilities for each pair of positions across the interface. (**d**) Three general scenarios of mutations for two positions on opposite sides of the protein–protein interface. Only the first one describes correlated mutations between the positions, with a MI of 1. For the latter two scenarios, the MI is 0 due to a lack of correlation or a lack of mutation, respectively.

from physical interaction. Thus, we set out to investigate from so-called 'first principles' the degree to which the energetic stability of a natural protein–protein complex underlies the phenomenon of the co-adaptability of its interface. Furthermore, our method enables us to focus on those pairs that demonstrate strong potential for co-adaptivity beyond a naive consideration of all pairs within spatial proximity. On the other hand, since our analysis is performed using a fixed-backbone procedure and enumeration of pair-wise adaptivity (Section 2), we do not expect to see many pairs of co-adapting residues that are distant in the native structure.

We provide an unbiased view of a dataset of >200 protein complexes consisting of both obligate and transient heterodimeric complexes. In brief, we adopt a structural and physical energy-based model, for which the lowest energy amino acid sequences (and their side chain rotamers) occur with higher probability. In the past, this model has been successfully employed to detect both the top 100 sequence predictions for protein design (Fromer and Yanover, 2009; Fromer *et al*., 2010) and also the probability of assigning particular amino acids at any residue (Fromer and Shifman, 2009). This approach bypasses the need for evolutionary sequence alignments (Kuipers *et al*., 2009); instead, the detailed 3D structure for each complex is used to predict the Boltzmann distribution-weighted pair-wise amino acid probabilities without sequence alignment-based estimation of CM. As a measure of CM, we utilize the well-known mutual information (MI) score (Dunn *et al*., 2008; Halperin *et al*., 2006), which quantifies to what degree the amino acid identities at a specific residue increase our knowledge regarding the amino acids at another residue in the partner protein. We measured the MI for all pairs of positions in all protein–protein interfaces, and we find that naturally transient complexes possess greater relative capacity for CM than naturally obligate complexes when modeled identically. In addition, we analyzed a set of >50 homodimeric complexes and find that, despite having interface sizes similar to the obligate complexes, they exhibit a large co-adaptive potential similar to that measured for transient complexes.

2 METHODS

2.1 Datasets

Starting from the 327 non-redundant transient and obligate heterodimers in the manually curated set of (Mintseris *et al*., 2007), we removed all obsolete and retracted PDB entries. Considering only cases for which the dimers had exactly two protein chains, we calculated the interface of the complex to be those positions for which some atom is within 4 Å of an atom in the other monomer. We removed from our analysis those complexes with >100 residues on both sides of the interface, yielding 210 heterodimers in total: 136 transient and 74 obligate complexes. Homodimers were retrieved from the collection of 76 non-redundant homodimers (Ponstingl *et al*., 2000). Applying the same filtration steps as above, we obtained a set of 53 homodimers. The list of all 263 dimeric complexes and the protein chains analyzed can be found at: http://www.protonet.cs.huji.ac.il/obligate_transient/complexes.txt.

2.2 Atomic modeling and calculation of probabilities

Each complex was modeled in one of two forms, either as (i) the native structure of the dimeric complex or as (ii) two unbound monomers. To obtain the unbound monomers, we simply separated the two sides of the dimer so that they would not interact, but implicitly assumed that there are no conformational changes that occur upon dissociation; this artificial assumption was critical in providing an unbiased comparison between the obligate and transient complexes since the obligate complexes cannot naturally exist as unbound proteins. We then used these two states to compose the modeling scenarios of interest. The 'one-state' (obligate-like) scenario consisted of a single state of the dimer (Fig. 1b), whereas the 'two-state' (transient-like) scenario consisted of two uniformly weighted states: the native dimer and the unbound monomers (Fig. 1a). Note that a uniform weighting was chosen to model all complexes under the same hypothetical two-state condition. For each structure, pair-wise energy calculations were performed using the widely used Rosetta package for protein design (Kuhlman and Baker, 2000), with a backbone-dependent rotamer library (Dunbrack and Karplus, 1993). The interface positions were allowed to mutate to all amino acids except cysteine, while all other positions in the proteins were held fixed to their native amino acid identities and

conformations, though the interactions of these fixed residues were also included in these calculations. This simplistic approach was adopted in order to accommodate the unbiased modeling of all sizes of complexes (ranging from 12 to 100 interfacial residues). For each design scenario (one- or two-state), a corresponding probabilistic graphical model was constructed (Fig. 1c), where nodes correspond to interface residues and the edges between them consist of the Rosetta interactions between residues that assume particular rotameric states. We employed the belief propagation algorithm in order to calculate all pair-wise amino acid–amino acid probabilities for each pair of interface positions in the two structures. The algorithm was run until convergence, with a limit of 10^6 messages passed before termination. When modeling the homodimeric complexes, we 'linked' each pair of corresponding positions in the two monomers so that their amino acids would necessarily be chosen consistently. See (Fromer and Yanover, 2009; Fromer *et al*., 2010) for full details on this algorithm for modeling proteins with multiple structural states and the benchmarking of its success.

2.3 MI score

As formulated in Mintseris *et al*. (2007), the MI for a pair of protein positions is the information-theoretic relative entropy between the predicted pair-wise distribution of amino acids and the product of their marginal probabilities. That is, for position i in monomer A and position j in monomer B, their MI is:

$$\mathrm{MI} = \sum_{A_i}^{20aa} \sum_{B_j}^{20aa} P(A_i, B_j) \log_2 \frac{P(A_i, B_j)}{P(A_i)P(B_j)}$$

where $P(A_i, B_j)$ denotes the probability of amino acids A_i and B_j at those positions. The MI is bounded from below by 0, and the larger the value, the more the two positions show inter-dependence in their amino acid choices.

For the belief propagation algorithm runs described above, we used a stochastic noise level (temperature) of 0.51, which we have previously found to yield reasonable results (Fromer *et al*., 2010). As control simulations, we used a very high temperature of 1.85, in order to obtain a per-complex estimation of the background level of noisy MI values. For each complex, we required that a significant MI value be within the 3% highest values of the control MI distribution for that complex modeled in the two-state scenario. Pairs of positions with significant MI values were those designated as having a CM between them. Note that other temperatures and percentage cutoffs yielded results quite similar to those shown here. The number of CM predicted for each complex in each scenario can be found at http://www.protonet.cs.huji.ac.il/obligate_transient/CM_ratios.txt.

2.4 Evolutionary profiles

Evolutionary profiles were extracted from the HSSP (homology-derived secondary structure of proteins) database, as described previously (Fromer *et al*., 2010) and plotted using the WebLogo application (weblogo.berkeley.edu).

3 RESULTS

For each particular protein–protein complex, we modeled it in two alternative scenarios. In the two-state scenario (Fig. 1a), the two proteins can either be in their native dimeric complex or as unbound monomers, where these two states are weighted uniformly. In the one-state scenario (Fig. 1b), the two proteins are only allowed to exist in a single state as a bound dimer. The motivation behind these alternative scenarios is to model 'obligate-like' behavior as one-state, where the two proteins are always in complex; on the other hand, 'transient-like' behavior requires that the proteins be physically stable both in complex and as isolated monomers. Our goal was to analyze the consequences of modeling each complex in both the one- and two-state scenarios and compare these analyses between the obligate and transient complexes.

To model the interface of a particular pair of proteins in either scenario, the structures of all states are simulated using a probabilistic graphical model. In the model (Fig. 1c), nodes correspond to interface residues and the edges between them consist of pseudo-physical atomic interaction energies. We employ the belief propagation algorithm to calculate all pair-wise amino acid–amino probabilities (Fromer and Yanover, 2008; Fromer *et al*., 2010), from which we determine if each particular pair exhibits CM (Fig. 1d) by calculating the MI score, where higher MI values correspond to larger degrees of correlated variability between the positions (for full details, see Section 2).

As was previously observed by us (Fromer and Yanover, 2009; Fromer *et al*., 2010) and others (Mandell and Kortemme, 2009), there are many cases for which standard, off-the-shelf all-purpose protein design calculations fail. Indeed, we found here that the design procedure often over-predicts certain amino acids, such as serine (S), threonine (T) and glycine (G). Nevertheless, the unbiased simulation of over 200 protein complexes comprising >150 000 designed inter-molecular pairs of positions allowed us to overcome this shortcoming, obtaining an abundance of high level information that is both consistent and statistically significant.

In our analysis, we studied a total of 210 non-redundant heterodimeric complexes: 136 naturally transient complexes and 74 naturally obligate complexes. For each such complex, the interface positions were defined as those in contact with positions in the opposing protein. For illustration, we consider PDB code 1KZY as an example of a transient complex (Fig. 2). 1KZY is the naturally transient heteromer of tumor suppressor p53 complexed with the tandem BRCT region of p53-binding protein 1 (53BP1) (Joo *et al*., 2002). The 3D structure is composed of two p53 chains and two 53BP1 chains. We modeled the interface between one of each of these protein chains (1KZY, chains A and C). This complex is of special interest as residues of the p53 DNA-binding surface are often mutated in cancer cells. Past analysis of experimental mutations has contributed to the estimation of the conservation levels and features of the interface (Joo *et al*., 2002).

Figure 2 shows the evolutionary profile (HSSP) for each of the positions in 1KZY, as well as the designed profiles in the one- and two-state scenarios. The number of CM increases from the modeling of the two-state scenario to the one-state scenario (from 6 to 10). Specifically, we observe that CM pairs found in the two-state scenario are almost always present in the one-state scenario as well, albeit with higher MI values (gray arrows). And, there are new CM not observed in the two-state scenario, which appear only in the one-state scenario (blue arrows). In addition, several general trends for the interface residues can be drawn from this example and others: (i) sequences predicted by computational fixed-backbone design are only moderately similar to natural sequences, with sequence identity remaining at ∼30% (Kuhlman and Baker, 2000); (ii) furthermore, the evolutionary profile is often significantly less diverse than the mutational profiles derived from the computational design approach, with only a few of the designed interface positions remaining highly conserved; (iii) an interface that is naturally rich in charged residues often retains its charged nature, although the positions at which specific amino acids appear are not always conserved; (iv) the number of glycine and serine, and to a lesser extent threonine, tends to increase relative to the HSSP profile

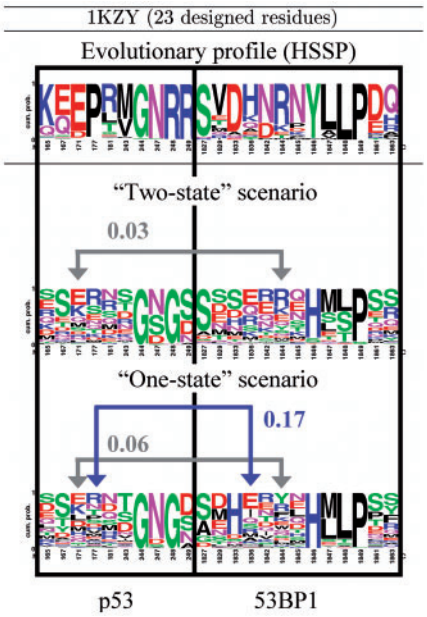

Fig. 2. Case study examples of CMs in the protein–protein interface of the p53–53BP1 complex. The total number of residues modeled and the relevant PDB identifiers are as indicated. The evolutionary positional sequence profiles (HSSP), as well as the predicted profiles for the two- and one-state scenarios, are shown. Representative examples of the predicted inter-molecular CM are demonstrated, where CM present in both the two- and one-state scenarios are depicted in gray, and those unique to the one-state scenario are depicted in blue. In this example, there are six CM pairs in the two-state scenario and 10 CM pairs in the one-state scenario.

(see examples, Supplementary Fig. S1). Note that similar trends, albeit with different intensities, were found in many of the 210 complexes studied here, irrespective of the number of interface residues that were included in the design procedure.

As was similarly observed in Mintseris and Weng (2003), we find that the size of the interface in obligate dimers is significantly greater than in transient dimers (Fig. 3a). As illustrated (Fig. 2), we modeled the interface of each such complex in both the two- and one-state scenarios and counted the number of significant CM predicted in each scenario (Fig. 3b). To determine which CMs are significant for each complex, we considered only those with MI values in the high end of the distribution for a control simulation (see Section 2). Not surprisingly, due to the larger number of pairs of positions available to be potentially correlated in the naturally obligate complexes (Fig. 3b, bottom), they exhibit larger absolute numbers of CM than transient complexes (top). Also, for each category of natural complex, it was expected that there would be significant increases in the number of CM between the two- and one-state models of the same complex (Fig. 3b, rightward shift of gray bars to black bars). However, as we will see below, the degree of this shift is significantly different for the naturally transient and naturally obligate complexes.

Since not all possible pairs of positions across a particular interface are likely to have physically induced CM between them (e.g. distant residues), we accounted for this by considering, for each complex, only the number of pairs falling within the maximal distance for which any significant CM was observed (Fig. 3c). We denoted the size of this smaller set of pairs as the 'effective interface size'. Still, these effective interface sizes show the same trend of larger interfaces for the obligate complexes than for the transient complexes. Next, we used these effective interface sizes to normalize the absolute numbers of CM predicted for each complex. When analyzing these normalized CM rates, the statistical differences between the transient and obligate complexes were eliminated (Fig. 3d). That is, when considering a protein in either of the two- (gray) or one-state (black) scenarios, the distinction between the naturally transient (top) and naturally obligate (bottom) complexes was insignificant (unpaired t-test, $P = 0.26$ and 0.50, for the two-state and one-state models, respectively). One of the main results from Mintseris and colleagues (Mintseris and Weng, 2005) is that obligate interfaces exhibit much higher degrees of CM than transient interfaces. Here, we also find there to be a large gap between the (normalized) number of CM for transient complexes modeled in the two-state scenario (Fig. 3d, top, gray bars) and for obligate complexes modeled in the one-state scenario (Fig. 3d, bottom, black bars). This difference, between obligate and transient complexes modeled in their natural scenarios (i.e. obligate complex in one-state model, and transient complex in two-state model), is extremely significant ($P < 10^{-30}$). Expectedly, this trend is even stronger when comparing absolute numbers of CM (Fig. 3b).

Overall, we conclude that our modeling of the two- and one-state scenarios for both transient and obligate complexes is quite similar (Fig. 3d) and thus unbiased. Nevertheless, when we calculated the ratio of CM in the one-state scenario to the CM in the two-state scenario, individually for each complex, and compared the transient and obligate datasets (Fig. 4a), we found the transient complexes to have significantly greater increases in CM ($P = 0.004$). From this result, we postulate that the interfaces of naturally transient complexes have a substantial but untapped potential for compensatory mutations and adaptability. This capacity could be activated if the partner proteins were permitted to be 'unnaturally' highly optimized for their dimeric interaction, to the exclusion of their monomeric forms (and possibly other interactions).

To test whether this observation is indeed a property of the interface, or the byproduct of a genuine difference in the stability and adaptability of the proteins as a whole, we performed the same calculations for pairs of positions on the same side of each interface (intra-molecular CM, Fig. 4b). For the intra-molecular CM, we found only a much lesser degree of difference in this ratio for the transient and obligate complexes ($P = 0.054$). On the one hand, this may be expected since the interface positions were specifically chosen as those capable of inter-molecular interaction; nonetheless, since these ratios are normalized by the number of CM in the two-state scenarios (also expected to be lower for intra-molecular pairs), this phenomenon cannot easily be explained away.

Next, we considered a set of 53 non-redundant homodimeric proteins and subjected them to the same simulations, of either two- or one-state scenarios. Note that we constructed these protein design calculations such that each position in each monomer would choose only the same amino acid as the corresponding position in the second monomer, thus properly simulating a homodimeric protein. Table 1 shows a summary of the results for all three categories of complexes. We see that the homodimers have interface sizes significantly larger than those of the transient complexes, but rather similar to the naturally obligate complexes. The same is true for

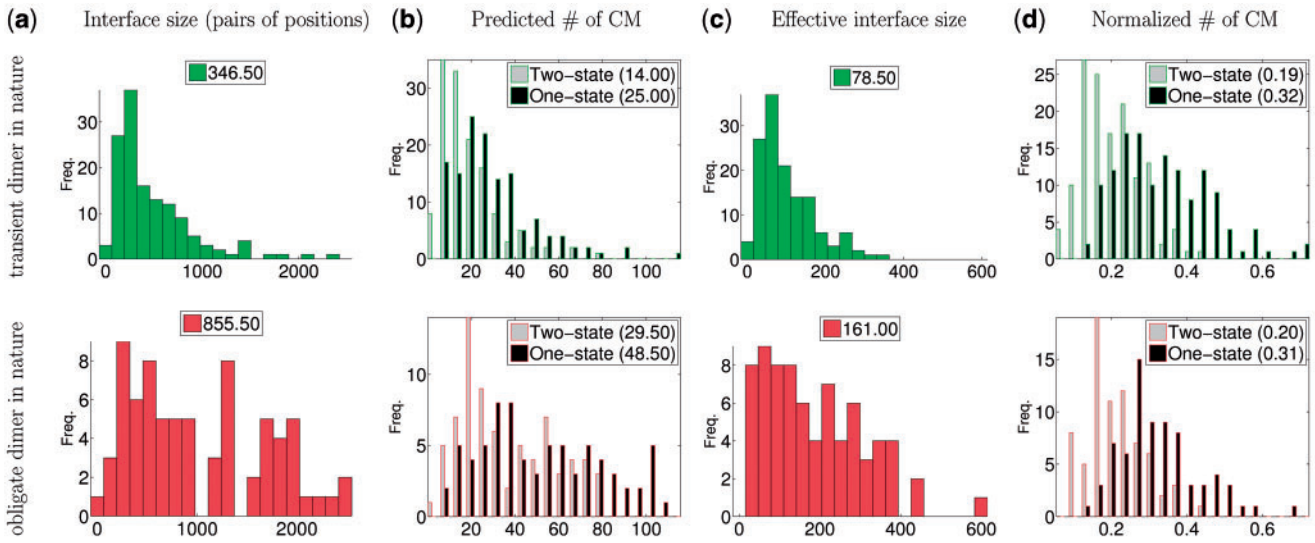

Fig. 3. Interface size and number of predicted CM for transient and obligate complexes. (**a**) For each complex, the size of the interface is the product of the number of residues determined to be interacting on each side of the heterodimer. (**b**) Histogram of the numbers of predicted CM for transient (top) and obligate (bottom) complexes, under two- (gray bars) and one-state (black bars) modeling. (**c**) Effective interface sizes, after subtracting pairs of inter-molecular residues more distant than those observed to have some CM in that complex. (**d**) Frequencies of CM after normalizing by the effective interface size for each complex, respectively. For all plots, numbers in the legends indicate the median values of the distributions.

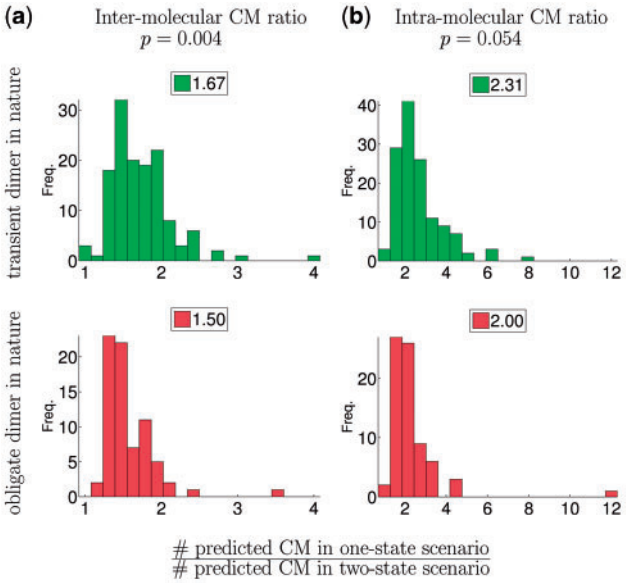

Fig. 4. Ratios of the number of predicted CM in the one-state model to the number of CM in the two-state model. (**a**) For each complex, the ratio of inter-molecular CM. The distributions for the obligate and transient complexes are significantly different, though this is not evident from the overall CM distributions in the two- or one-state models analyzed without respect to the particular complex from which they are derived (Fig. 3d). (**b**) For each complex, the ratio of intra-molecular CM, i.e. those CM between a pair of interface positions on the same side of the interface. For all plots, numbers in the legends indicate the median distribution values.

Table 1. Median values of interface sizes, number of predicted CM and CM ratios between the one- and two-state scenarios for different complex types

Type[a]	Number of complexes	Interface size (*P*-value)	CM in two-state scenario (*P*-value)	CM in one-state scenario (*P*-value)	Per-complex CM ratio (*P*-value)
T	136	**346.5**	**14.0**	**25.0**	1.67 (0.004)
O	74	855.5 (10^{-11})	29.5 (10^{-11})	48.5 (10^{-11})	**1.50**
H	53	702.0 (10^{-29})	28.0 (10^{-5})	46.0 (10^{-7})	1.70 (0.018)

[a]The types of complexes include transient (T) and obligate (O) heterodimers and homodimers (H). Values in bold indicate the statistical outlier in each column, and numbers in parentheses denote the *P*-values between the corresponding distribution and the outlier distribution.

the absolute numbers of CM when modeled in either the two- or one-state scenarios, where the homodimers resemble the obligate heterodimers. Nevertheless, when considering the ratios between CM in the one-state scenario to the CM in the two-state scenario, the homodimers are different from the obligate complexes and far more similar to the transient ones. Thus, both the transient heterodimers and the homodimers have a similarly large capacity for co-adaptibility of their interfaces, which is intrinsically lacking in the structures of the obligate heterodimers. Of interest, the surprisingly strong CM potential in homodimers could be explained simply by the structural identity of the two monomers, as was previously argued regarding the self-affinity of protein structures (Pereira-Leal *et al.*, 2007).

4 DISCUSSION

The main finding of this article is that the interaction interface of obligate protein–protein heterodimers has significantly less potential for co-optimization across the complex than that of transient heterodimers (and homodimers). We hypothesize that this may derive from the fact that evolutionary processes have selected for transient interfaces with structures that have consistently maintained their ability to adapt as new partners are added to their binding repertoire (and existing partners are removed). Indeed, the energetic consideration of protein interfaces in determining their specificity (Carbonell *et al*., 2009; Fromer and Shifman, 2009; Fromer and Yanover, 2009) has provided insight into the participation of these proteins in rich interactions with multiple partners and in networks in general (Tyagi *et al*., 2009).

We propose that the 'unused' potential for compensatory mutations observed in transient interfaces does not derive solely from the fact that these two proteins are not always found together (Tyagi *et al*., 2009). If this were the case, then we would expect that modeling a transient complex in the one-state scenario would not yield a larger relative increase in CM than when performing the same procedure for obligate complexes (Fig. 4a). Thus, we hypothesize that the increased concerted capacity for adaptation present in transient dimers is an intrinsic requirement of the structures of such interfaces—e.g. so that they possess the means to dynamically adapt to new interactions. This greater potential for pair-wise adaptability across transient interfaces is also consistent with previous research, where it was found that transient complexes have greater optimization for electrostatic stability (charge pairing) across the interface (Brock *et al*., 2007).

Moreover, we have shown that the properties of the interfaces of obligate and transient complexes do not extend to the entire proteins. We found that other properties of the complexes, such as intra-molecular CM modeling (Fig. 4b), the absolute size of the protein dimers, and the number of additional chains present in the solved 3D structures are not correlated with the properties of the 'coadaptive potential'. Previously, in a study on protein complexes resulting from duplication events in yeast, CM signals could not be detected (Hakes *et al*., 2007). On the other hand, we argue that by using a protein set that covers >260 complexes (including homodimers) from a broad range of environmental and phylogenetic contexts, we recover a statistically robust signal of potential biophysical CM for protein interfaces (Table 1).

Here, we also performed a study of protein homodimers (obligate and transient). While the global interface properties (size) resemble the obligate set, the CM properties resemble that of protein complexes that are transient in nature. The evolution of a homodimer is often associated with a large and extended interface, mainly for the purpose of geometric complementarity (Lukatsky *et al*., 2007). Therefore, we find it interesting that the co-adaptive potential is still strong enough for multiple mutations to occur (experimentally or by computational design) without forfeiting energetic stability (Table 1).

In summary, we note that several high-level principles have been proposed for the evolution of structurally functional protein complexes: those that are obligate or transient (this study), having multiple partners (Fromer and Shifman, 2009; Fromer and Yanover, 2009; Humphris and Kortemme, 2007) and maintaining partners from a similar superfamily or fold (Lukatsky *et al*., 2007). It is an intriguing possibility that the 'transient-like' nature of homodimers and their predicted adaptability to mutations may also underlie the evolution of heterodimers that share the same fold (Pereira-Leal *et al*., 2007).

Funding: EU Framework VII Prospects consortium; ISF 592/07; Sudarsky Center for Computational Biology (to M.F.).

Conflict of Interest: none declared.

REFERENCES

Ansari,S. and Helms,V. (2005) Statistical analysis of predominantly transient protein-protein interfaces. *Proteins*, **61**, 344–355.

Berezovsky,I.N. *et al*. (2007) Positive and negative design in stability and thermal adaptation of natural proteins. *PLoS Comput. Biol.*, **3**, e52.

Brock,K. *et al*. (2007) Optimization of electrostatic interactions in protein-protein complexes. *Biophys. J.*, **93**, 3340–3352.

Capra,J.A. and Singh,M. (2008) Characterization and prediction of residues determining protein functional specificity. *Bioinformatics*, **24**, 1473–1480.

Carbonell,P. *et al*. (2009) Energetic determinants of protein binding specificity: insights into protein interaction networks. *Proteomics*, **9**, 1744–1753.

Chi,C.N. *et al*. (2008) Reassessing a sparse energetic network within a single protein domain. *Proc. Natl Acad. Sci. USA*, **105**, 4679–4684.

Deeds,E.J. *et al*. (2007) Robust protein-protein interactions in crowded cellular environments. *Proc. Natl Acad. Sci. USA*, **104**, 14952–14957.

Dunbrack,R.L., Jr. and Karplus,M. (1993) Backbone-dependent rotamer library for proteins. Application to side-chain prediction. *J. Mol. Biol.*, **230**, 543–574.

Dunn,S.D. *et al*. (2008) Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. *Bioinformatics*, **24**, 333–340.

Fariselli,P. *et al*. (2001) Progress in predicting inter-residue contacts of proteins with neural networks and correlated mutations. *Proteins*, **45**(Suppl. 5), 157–162.

Fromer,M. and Shifman,J.M. (2009) Tradeoff between stability and multispecificity in the design of promiscuous proteins. *PLoS Comput. Biol.*, **5**, e1000627.

Fromer,M. and Yanover,C. (2008) A computational framework to empower probabilistic protein design. *Bioinformatics*, **24**, i214–i222.

Fromer,M. and Yanover,C. (2009) Accurate prediction for atomic-level protein design and its application in diversifying the near-optimal sequence space. *Proteins*, **75**, 682–705.

Fromer,M. *et al*. (2010) Design of multispecific protein sequences using probabilistic graphical modeling. *Proteins*, **78**, 530–547.

Hakes,L. *et al*. (2007) Specificity in protein interactions and its relationship with sequence diversity and coevolution. *Proc. Natl Acad. Sci. USA*, **104**, 7999–8004.

Halperin,I. *et al*. (2006) Correlated mutations: advances and limitations. A study on fusion proteins and on the Cohesin-Dockerin families. *Proteins*, **63**, 832–845.

Humphris,E.L. and Kortemme,T. (2007) Design of multi-specificity in protein interfaces. *PLoS Comput. Biol.*, **3**, e164.

Joo,W.S. *et al*. (2002) Structure of the 53BP1 BRCT region bound to p53 and its comparison to the Brca1 BRCT structure. *Genes Dev.*, **16**, 583–593.

Jothi,R. *et al*. (2006) Co-evolutionary analysis of domains in interacting proteins reveals insights into domain-domain interactions mediating protein-protein interactions. *J. Mol. Biol.*, **362**, 861–875.

Kann,M.G. *et al*. (2009) Correlated evolution of interacting proteins: looking behind the mirrortree. *J. Mol. Biol.*, **385**, 91–98.

Kuhlman,B. and Baker,D. (2000) Native protein sequences are close to optimal for their structures. *Proc. Natl Acad. Sci. USA*, **97**, 10383–10388.

Kuipers,R.K. *et al*. (2009) Correlated mutation analyses on super-family alignments reveal functionally important residues. *Proteins*, **76**, 608–616.

Kundrotas,P.J. and Alexov,E.G. (2006) Predicting residue contacts using pragmatic correlated mutations method: reducing the false positives. *BMC Bioinformatics*, **7**, 503.

Lee,B.C. and Kim,D. (2009) A new method for revealing correlated mutations under the structural and functional constraints in proteins. *Bioinformatics*, **25**, 2506–2513.

Lukatsky,D.B. *et al*. (2007) Structural similarity enhances interaction propensity of proteins. *J. Mol. Biol.*, **365**, 1596–1606.

Madaoui,H. and Guerois,R. (2008) Coevolution at protein complex interfaces can be detected by the complementarity trace with important impact for predictive docking. *Proc. Natl Acad. Sci. USA*, **105**, 7708–7713.

Mandell,D.J. and Kortemme,T. (2009) Backbone flexibility in computational protein design. *Curr. Opin. Biotechnol.*, **20**, 420–428.
Mintseris,J. *et al*. (2007) Integrating statistical pair potentials into protein complex prediction. *Proteins*, **69**, 511–520.
Mintseris,J. and Weng,Z. (2003) Atomic contact vectors in protein-protein recognition. *Proteins*, **53**, 629–639.
Mintseris,J. and Weng,Z. (2005) Structure, function, and evolution of transient and obligate protein–protein interactions. *Proc. Natl Acad. Sci. USA*, **102**, 10930.
Ofran,Y. and Rost,B. (2003) Analysing six types of protein-protein interfaces. *J. Mol. Biol.*, **325**, 377–387.
Oliveira,L. *et al*. (2002) Correlated mutation analyses on very large sequence families. *Chembiochem*, **3**, 1010–1017.
Pazos,F. *et al*. (1997) Correlated mutations contain information about protein-protein interaction. *J. Mol. Biol.*, **271**, 511–523.
Pazos,F. and Valencia,A. (2008) Protein co-evolution, co-adaptation and interactions. *EMBO J.*, **27**, 2648–2655.
Pereira-Leal,J.B. *et al*. (2007) Evolution of protein complexes by duplication of homomeric interactions. *Genome Biol.*, **8**, R51.
Ponstingl,H. *et al*. (2000) Discriminating between homodimeric and monomeric proteins in the crystalline state. *Proteins*, **41**, 47–57.
Shackelford,G. and Karplus,K. (2007) Contact prediction using mutual information and neural nets. *Proteins*, **69** (Suppl. 8), 159–164.
Smith,G.R. and Sternberg,M.J. (2002) Prediction of protein-protein interactions by docking methods. *Curr. Opin. Struct. Biol.*, **12**, 28–35.
Thomas,J. *et al*. (2009) Graphical models of protein-protein interaction specificity from correlated mutations and interaction data. *Proteins*, **76**, 911–929.
Tyagi,M. *et al*. (2009) Exploring functional roles of multibinding protein interfaces. *Protein Sci.*, **18**, 1674–1683.
Weigt,M. *et al*. (2009) Identification of direct residue contacts in protein-protein interaction by message passing. *Proc. Natl Acad. Sci. USA*, **106**, 67–72.
Yeang,C.H. and Haussler,D. (2007) Detecting coevolution in and among protein domains. *PLoS Comput. Biol.*, **3**, e211.