Evolutionary constraints on structural similarity in orthologs and paralogs

Mark E. Peterson,[1,2,3]* Feng Chen,[4,5] Jeffery G. Saven,[4] David S. Roos,[5] Patricia C. Babbitt,[1,2,3] and Andrej Sali[1,2,3]*

[1]Department of Bioengineering and Therapeutic Sciences, University of California, San Francisco, San Francisco, California 94158

[2]Department of Pharmaceutical Chemistry, University of California, San Francisco, San Francisco, California 94158

[3]California Institute for Quantitative Biosciences, University of California, San Francisco, San Francisco, California 94158

[4]Department of Chemistry, University of Pennsylvania, Philadelphia, PA 19104

[5]Department of Biology and Penn Genomics Institute, University of Pennsylvania, Philadelphia, PA 19104

Received 9 December 2008; Revised 29 March 2009; Accepted 30 March 2009
DOI: 10.1002/pro.143
Published online 16 April 2009 proteinscience.org

Abstract: Although a quantitative relationship between sequence similarity and structural similarity has long been established, little is known about the impact of orthology on the relationship between protein sequence and structure. Among homologs, orthologs (derived by speciation) more frequently have similar functions than paralogs (derived by duplication). Here, we hypothesize that an orthologous pair will tend to exhibit greater structural similarity than a paralogous pair at the same level of sequence similarity. To test this hypothesis, we used 284,459 pairwise structure-based alignments of 12,634 unique domains from SCOP as well as orthology and paralogy assignments from OrthoMCL DB. We divided the comparisons by sequence identity and determined whether the sequence-structure relationship differed between the orthologs and paralogs. We found that at levels of sequence identity between 30 and 70%, orthologous domain pairs indeed tend to be significantly more structurally similar than paralogous pairs at the same level of sequence identity. An even larger difference is found when comparing ligand binding residues instead of whole domains. These differences between orthologs and paralogs are expected to be useful for selecting template structures in comparative modeling and target proteins in structural genomics.

Keywords: orthology; paralogy; sequence-structure relationship; orthologs prediction; homology modeling

Introduction

One of the foundations of molecular biology is that a protein's sequence determines its structure, which in turn determines how the protein functions. These sequence–structure–function dependencies allow us to better deduce evolutionary relationships between proteins and between organisms, to better discern the functions of the thousands of genes from many genome sequencing projects, and to improve design of new protein functions and drugs.

Protein sequence–structure–function relationships have been investigated and quantified in various ways. Structural similarity between proteins (measured, for example, by the root mean square deviation (RMSD) between the backbone atoms of the common cores of two protein structures) is clearly related to their sequence similarity.[1–6] Other studies have established the level of sequence similarity at which structural similarity is likely to be observed.[7,8]

Additional Supporting Information may be found in the online version of this article.

Mark E. Peterson and Feng Chen contributed equally to this work.

Grant sponsor: The Sandler Family Supporting Foundation; NIH; Grant numbers: P01 GM71790, R01 GM60595, R01 GM61267, R01 GM54762, U54 GM074945.

*Correspondence to: Mark E. Peterson or Andrej Sali, Mission Bay, Byers Hall, 1700 4th Street, Suite 503B, University of California, San Francisco, CA 94158-2330. E-mail: markpete@salilab.org or sali@salilab.org

Quantitative analyses of the relationship between sequence similarity and functional similarity have frequently focused on the degree of sequence similarity required to be able to transfer functional annotations from one protein to another. Although not without limitations,[9] Enzyme Commission (E.C.) numbers[10] constitute one of the most common ways of specifying protein function for such studies. Sequence similarity thresholds above which E.C. numbers can be reliably transferred from one protein to another have been suggested.[6,11–13] Measures of function other than E.C. number have also been used,[14,15] and the expected functional similarity between pairs of proteins for a broad range of sequence identities has been determined.[12,16,17]

Relationships between structural similarity and functional similarity have also been studied.[6,16,18–20] A correlation has been observed between functional similarity and RMSD between pairs of proteins,[6] although structure and function generally seem to be less closely correlated than sequence and structure.[6,20]

Much of the power of the relationships between sequence, structure, and function lies in conclusions applying broadly to many classes of proteins. However, we can ask more specifically if the relationships are quantitatively different depending on the subset of proteins examined. To some extent, this approach has been taken by those who have examined sequence-structure-function relationships by protein family or by fold class. For example, it has been found that sequence similarity and structural similarity are correlated within protein families,[21] but that this relationship varies between different protein families.[22] In contrast, sequence-structure relationships do not appear to be significantly determined by fold class alone.[3,6]

We present another informative way of selecting subsets of proteins to examine: dividing proteins according to orthology and paralogy.[23] Orthologs are homologous proteins that are related by speciation events and tend to show more functional similarity than other homologs.[24,25] Paralogs are homologous proteins that are related by a gene duplication event, and tend to show less functional similarity than orthologs.[25] The impact of orthology on functional similarity has been studied in the past,[26,27] and we add to these studies by examining orthology and paralogy from a structural perspective. Surprisingly, little is known empirically about the impact of orthology on the relationship between protein sequence and structure. It is generally recognized that proteins with similar functions tend to have similar structures. Consider a particular reference protein and its paralogs and orthologs. A paralog, having relaxed functional constraints due to the possible redundancy of gene duplication, may be free to acquire sequence changes that alter its structure. An ortholog with the same sequence identity to the reference, however, must retain function and may share greater structural similarity with the reference protein than the paralog. Here, we test this hypothesis by quantitatively comparing the relationship between sequence and structure in orthologous and paralogous proteins. In so doing, we examine how similarity in functional constraints (as suggested by evolutionary relationship) impacts the relationship between sequence and structure.

The identification of orthologs and paralogs itself is not a solved problem.[25] Although it is impossible to reconstruct the ancestry of any particular gene with complete certainty, various methods have been developed for identifying orthology and paralogy and for assessing the accuracy of these identifications.[28] Classical strategies for identifying orthologs have used phylogenetic reconstruction[29–33] and typically involve reconciling gene and species trees. Various challenges with these approaches, including problems introduced by horizontal gene transfer, artifacts associated with the properties of trees, and computational expense, have led to the development of complementary approaches not requiring phylogenetic analysis.[34] The most commonly used databases of orthology and paralogy assignments, such as COG,[26,35] do not use tree reconciliation. Instead, a simplifying assumption is made that orthologous genes in pairs of genomes can be identified as symmetrical best hits when comparing gene sequences. For our study, we obtain orthology and paralogy assignments from the recently developed, automated ortholog identification tool, OrthoMCL DB,[36] which assigns orthology and paralogy using comparisons of full-length sequences, genome similarity, and Markov clustering. We chose to use OrthoMCL because in a recent performance evaluation,[28] OrthoMCL was shown to provide a good balance of sensitivity and specificity.

This study provides new insights into sequence–structure relationships, with implications for improving comparative modeling and structural genomics. Both of these applications employ sequence similarity as a predictor of structural similarity. Making such predictions more accurate by adding information about orthology should in turn allow better selection of templates in comparative modeling[37–41] and better selection of targets for structural genomics.[42–46]

We begin by presenting results from our large-scale analyses of whole protein domains and sets of ligand-binding residues within those protein domains. We find that orthologs do indeed tend to be more structurally similar than paralogs at the same level of sequence similarity (Results). We then present two examples taken from these large-scale analyses that demonstrate the overall trends found in the data, as well as two counterexamples. Next, we suggest an explanation for those results that depart from the more general trends (Discussion). Finally, we discuss possible applications of our work in comparative modeling and in structural genomics.

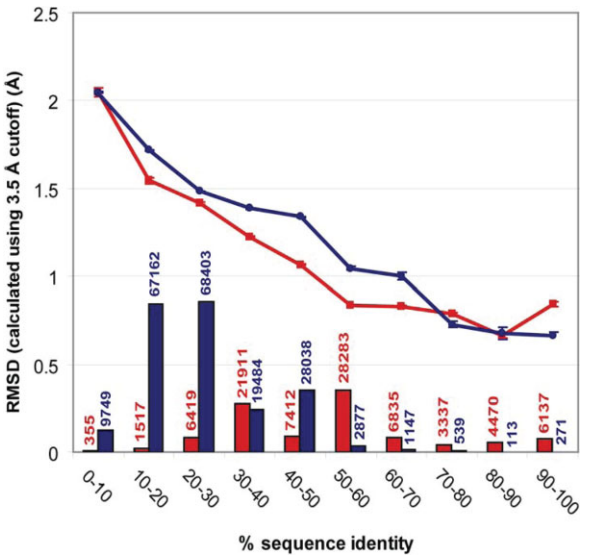

Figure 1. Global RMSD as a function of sequence identity for orthologous domain pairs (red squares) and paralogous domain pairs (blue triangles). RMSD calculated over alignment positions for which the C_α atoms from the aligned residues were within 3.5 Å of each other in the structural superposition. Larger values of RMSD indicate greater structural divergence. Error bars represent the 95% confidence interval of the mean RMSD for each sequence identity range, and reflect both the standard deviations of the RMSDs and the large sample sizes that were used. Sample sizes are shown for orthologous pairs (red bars) and paralogous pairs (blue bars) for each range of sequence identities.

Results

Comparisons of whole domains

We hypothesized that at a given sequence identity, structural similarity will be higher for orthologs than for paralogs. To quantify any such increase as a function of sequence identity, we plotted sequence identity *versus* structural divergence (measured by RMSD, the root-mean-square deviation between the aligned C_α atoms of two structures) separately for orthologous and paralogous domains (see Fig. 1). This plot includes 86,676 pairs of orthologs and 197,783 pairs of paralogs, and can be divided into two regions corresponding to sequence identities above and below 70% (Supporting information Table S1). We also constructed the corresponding plots using only crystallographically determined structures (all resolutions, resolutions better than 2.5 Å, and resolutions better than 2.0 Å); the resulting plots were similar to each other (data not shown).

The two curves have the largest separation for domains that share less than 70% sequence identity. In this range, orthologs are indeed substantially more structurally similar than paralogs of the same sequence identity (average C_α RMSDs of 1.05 ± 0.002 and 1.55 ± 0.001 Å, respectively; ranges given are 95% confidence intervals for the means). Because of the large sample sizes (72,732 and 196,860, respectively) in this range of sequence identity, the confidence intervals are small despite relatively large standard deviations for the C_α RMSDs (standard deviations are 0.30 and 0.29 Å for orthologous and paralogous pairs, respectively).

For pairs of domains with sequence identities above 70%, the observed differences between average C_α RMSDs for orthologs and paralogs essentially disappear: C_α RMSDs averaged 0.77 ± 0.01 and 0.72 ± 0.01 Å for orthologs and paralogs, respectively. Again, the confidence intervals are small because of the large sample sizes, although the sample sizes were smaller than those at less than 70% sequence identity (Fig. 1, Supporting information Table S1). Trends were similar when using both C_α and C_β atoms to calculate RMSDs, as well as when using all main-chain atoms (data not shown). We note that finding low RMSDs (below 1 Å) in this range of sequence identities is consistent with results from Chothia and Lesk[1] and Wilson *et al.*[6]

The effect of orthology versus paralogy on the relationship between sequence and structure may also be considered by specifying the level of structural similarity and comparing the sequence identities for orthologous and paralogous pairs. For example,

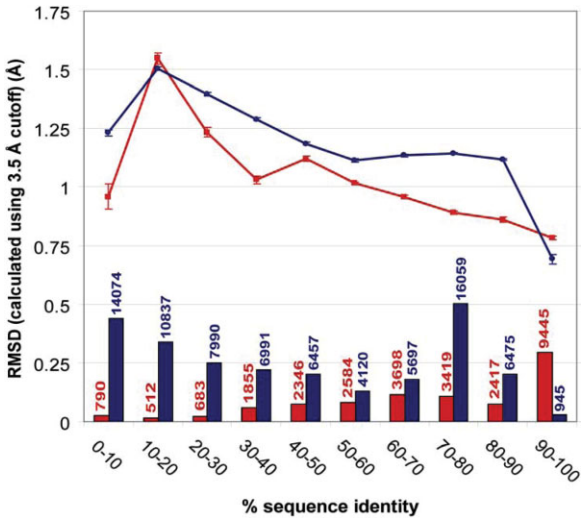

Figure 2. Local RMSD as a function of local sequence identity for orthologous domain pairs (red squares) and paralogous domain pairs (blue triangles). RMSD and sequence identity calculated over alignment positions for which one of the residues was designated a ligand-binding residue in LigBase and for which the C_α atoms from the aligned residues were within 3.5 Å of each other in the structural superposition. Larger values of RMSD indicate greater structural divergence. Error bars represent the 95% confidence interval of the mean RMSD for each sequence identity range. Sample sizes are shown for orthologous pairs (red bars) and paralogous pairs (blue bars) for each range of sequence identities.

interpolating between the points shown in Figure 1, at an average Cα RMSD of 1.0 Å, the corresponding sequence identities are 48% for orthologous pairs and 65% for paralogous pairs. This observation indicates that a paralogous pair of proteins has a sequence identity that is, on average, ~17 percentage points higher than a corresponding orthologous pair with the same structural similarity.

Comparisons of ligand-binding residues

Ligand-binding residues deliver function by providing specific interactions between proteins and their ligands (e.g., substrates, cofactors, other proteins, or inhibitors). As these residues tend to be conserved during evolution, we expect not only that the structural similarity between orthologs is greater than that for paralogs at any level of sequence similarity, but also that this difference in structural similarities should be larger for ligand-binding residues than for whole domains. Ligand-binding residues were identified using known complexed structures annotated in LigBase.[47] From our original sets of orthologous and paralogous domain pairs (described earlier), 5,066 and 28,938 pairs, respectively, had at least three aligned residue pairs that were identified as ligand-binding. The average number of aligned ligand-binding residues in a pair was 20. The plot of local C_α RMSD for these aligned ligand-binding residues versus local sequence identity (see Fig. 2) can be divided into three regions, corresponding to sequence identities below 20%, between 20 and 90%, and above 90% (Supporting information Table S2). Although the trends here are similar to those for whole-domain comparisons, we found, as expected, that there was a greater separation between the two curves over a larger range of sequence identities for ligand-binding residues than there was for whole domains.

Using a common reference structure to compare whole domains

We also compared orthologous versus paralogous sequence-structure relationships using a common query domain to limit each comparison to proteins of similar structure (e.g., in the same family). Specifically, we examined 3,816 triplets of proteins, each triplet consisting of a query domain, an ortholog of the query domain, and a paralog of the query domain. As before, orthology and paralogy assignments were obtained using OrthoMCL DB. Each point in Figure 3 shows the difference in sequence identities between the orthologous pair and the paralogous pair of the triplet, as well as the corresponding difference in structural similarities. If the type of evolutionary relationship between pairs of protein domains does not affect the relationship between sequence identity and structural similarity, then the trend line (fitted by linear least-squares regression) would be expected to intersect the origin, corresponding to equal sequence identities

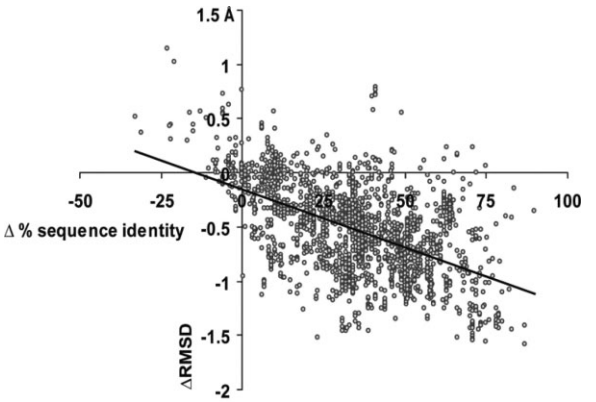

Figure 3. Each point represents a triplet of protein domains. A single triplet consists of a query domain, one ortholog to that query, and one paralog to that query. The *x*-axis shows the difference between the sequence identity of the ortholog to the query and the sequence identity of the paralog to the query. The *y*-axis shows the difference between the query-ortholog RMSD and the query-paralog RMSD. The trend line was fit by linear least-squares regression. The equation of the line is $y = -0.011x - 0.153$, and it intersects the *x*-axis at −14.2% (with a 95% confidence interval of (−14.86, −13.64)) and the *y*-axis at −0.15 Å (with a 95% confidence interval of (−0.180, −0.126)). Its R^2 value is 0.30, with the R^2 value representing a measure of its goodness-of-fit (the R^2 statistic can range from −1 to 1, with 1 representing perfect positive correlation and −1 representing perfect negative correlation).

resulting in equal structural similarities for both orthologs and paralogs. Instead, we observed a marked departure of the trend line from the origin. At the same level of structural similarity, paralogous domain pairs have sequence identities that are on average 10 percentage points higher than those of orthologous domain pairs. Similarly, for comparable sequence identities, orthologous domain pairs have Cα RMSDs that are 0.11 Å lower than those of paralogous domain pairs.

Investigations of representative and anomalous cases

We visually inspected a large number of cases, including both those that conformed to our hypothesis and those that did not. In Figure 4, we present two examples of using a common reference structure to compare whole domains: in each case both an ortholog and a paralog to the same domain were superposed on that domain's structure. The first case shows a relationship between domains of similar sequence identities that supports our hypothesis (27% sequence identity and 1.3 Å RMSD for the ortholog versus 30% sequence identity and 1.8 Å RMSD for the paralog) [Fig. 4(a)]. There are no large (≥15 residues) contiguous regions in which the residues of one of the homologs are closer to their aligned residues in the query

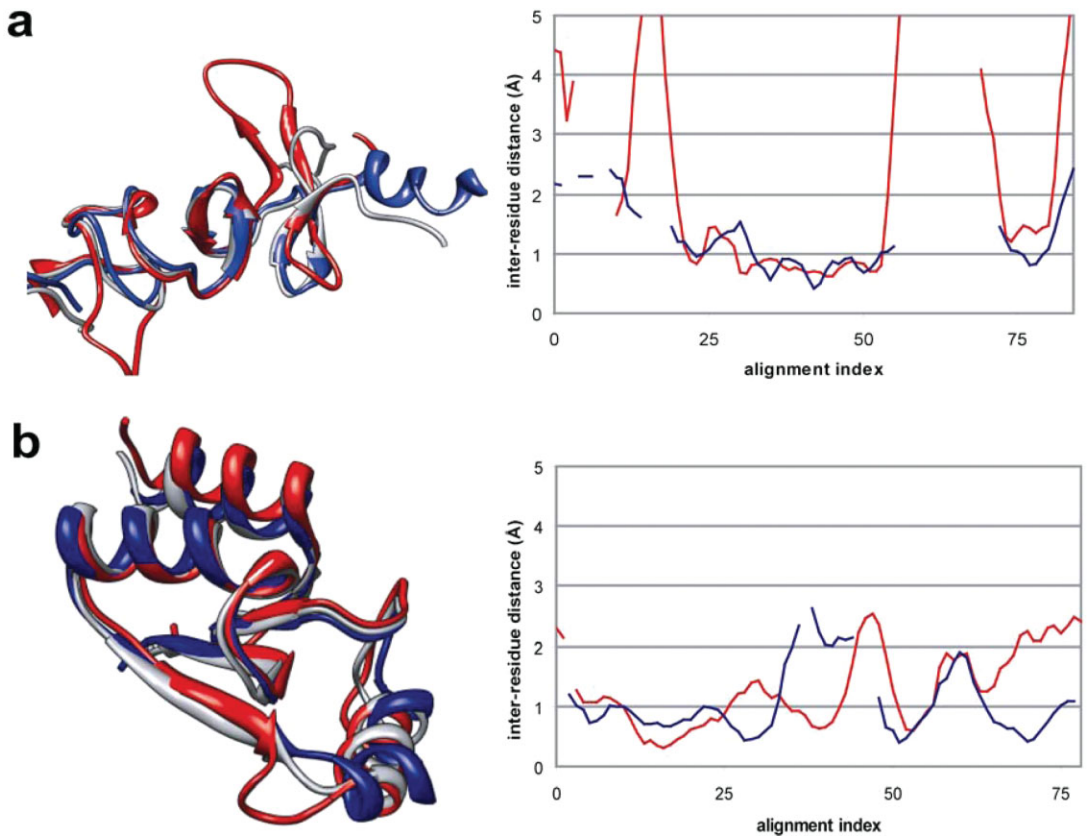

Figure 4. (a) Superposition of an ortholog (SCOP domain d1n8yc4, in red) and a paralog (SCOP domain d1igra3, in blue) onto middle domain of human Supernatant protein factor (SPF) (SCOP domain d1s78a3, in gray). The plot shows distances between middle domain of human receptor tyrosine-protein kinase erbB-2 and aligned C_α atoms in its ortholog (red curve) and its paralog (blue curve) after superposition onto erbB-2. Resolutions are listed for each domain beside their respective SCOP codes. (b) Superposition of an ortholog (SCOP domain d1q4jb2, in red) and a paralog (SCOP domain d4pgtb2, in blue) onto C-terminal domain of human Glutathione S-transferase (GST) (SCOP domain d1pkwa2, in gray). The plot shows distances between middle domain of human glutathione S-transferase (GST) and aligned C_α atoms in its ortholog (red curve) and its paralog (blue curve) after superposition onto glutathione S-transferase (GST) as a function of glutathione S-transferase (GST) residue number. Molecular graphics images were produced using the UCSF Chimera package[48] from the Resource for Biocomputing, Visualization, and Informatics at the University of California, San Francisco (supported by NIH P41 RR-01081).

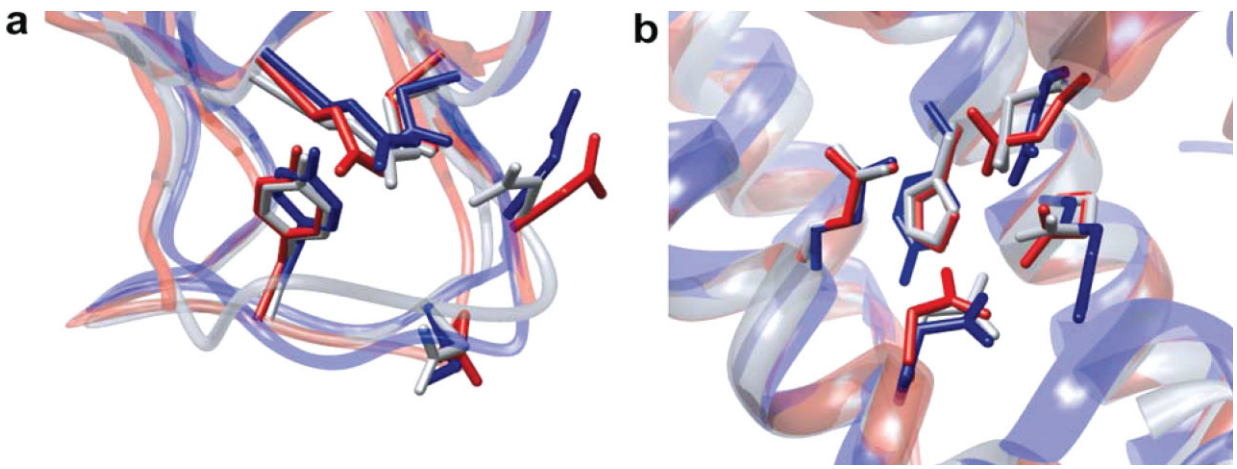

Figure 5. (a) Superposition of ligand-binding residues from an ortholog (SCOP domain d1nt0g1, in red) and a paralog (SCOP domain d1nzib1, in blue) onto those of human MBL-associated protein 19 (Map 19) (SCOP domain d1szba1, in gray). (b) Superposition of ligand-binding residues from an ortholog (SCOP domain d1xsm__, in red) and a paralog (SCOP domain d1smsb_, in blue) onto those of ribonucleotide reductase from *S. cerevisiae* (SCOP domain d1jk0a_, in gray).

than are those in the other homolog. There are also no contiguous regions longer than four residues in which one homolog is at least 0.5 Å closer to the query than the other homolog. Rather, the ortholog is more consistently closer to the query, but by a smaller margin than is seen for the second case below [Fig. 4(b)]. The higher structural similarity between the orthologs in this case is accompanied by greater functional similarity. Both of the orthologs form essential heterodimeric components of a neuregulin–receptor complex, while the paralog to the query domain forms a tetramer that binds insulin-like growth factor 1 (IGF1) with a high affinity and IGF2 with a lower affinity. Additional inspected cases consistent with our hypothesis displayed similar behavior; others showed more variation in the structural divergence between the query domain and its ortholog/paralog.

The second case [Fig. 4(b)] shows a counterexample (32% sequence identity and 1.5 Å RMSD for the ortholog versus 24% sequence identity and 1.2 Å RMSD for the paralog) in which the ortholog is more structurally divergent despite its higher sequence identity. In this counterexample, there is one region (alignment positions 35–45) in which the superposed ortholog is closer (0.55 Å on average) to the query structure than is the paralog. However, there are also two regions (alignment positions 28–34 and 62–78) in which the paralog is closer (1.2 and 1.1 Å on average) to the query structure than is the ortholog. Additional inspected cases inconsistent with our hypothesis displayed similar behavior; others showed less variation in the structural divergence from the query domain.

In Figure 5, we show examples of relationships similar to those aforementioned (in support of and against the hypothesis) for ligand-binding residues in particular. Over the sets of ligand-binding residues, the ortholog and paralog in Figure 5(a) have local RMSDs of 1.0 and 1.4 Å, respectively, to the ligand-binding residues of the query domain. RMSD values were calculated using all nonhydrogen atoms of the aligned residues, including the side chains. Those in Figure 5(b) have local RMSDs of 1.1 and 0.8 Å, respectively. The reference structures used for these cases share between 23 and 32% global sequence identity with the orthologs and paralogs shown. The atomic distances between the orthologs/paralogs and the query structures were quite variable (distances between aligned C_β atoms of ligand-binding residues varied between 0.1 and 1.4 Å). Both cases included positions at which the orthologs' ligand-binding residues were closer to and positions at which the ortholog's ligand-binding residues were more distant from the query than were the paralog's.

Discussion

Elaborating on previous work that examined sequence-structure relationships in proteins,[1–8,21,22] we asked whether orthologs at a particular level of sequence similarity show more structural similarity than paralogs at that same level of sequence similarity. To address this question, we identified pairs of orthologous and paralogous domains and compared the average structural divergences between pairs of orthologs and pairs of paralogs as a function of sequence identity. Our hypothesis was confirmed for sequence identities below 70%, with the greatest divergence between orthologs and paralogs seen when sequence identities were between 30 and 70%. We now discuss these results and their implications for comparative modeling and structural genomics.

Results by sequence identity range

The middle range of sequence identities (30–70%), in which our hypothesis was most strongly confirmed (average C_α RMSDs of 1.00 Å and 1.34 Å, for orthologs and paralogs, respectively), is also the range in which we expect our data to be most reliable. In this range, more reliable sequence alignments are possible than at lower sequence identities, and consequently more accurate estimates of the "true" sequence identity (that based on an alignment of evolutionarily equivalent positions) and RMSD are possible. Similarly, discrimination between orthologs and paralogs is also expected to be most accurate at this intermediate level of sequence similarity.

At very low sequence identities (below 30%), the structural differences between orthologs and paralogs (average C_α RMSDs of 1.47 and 1.63 Å, respectively), were somewhat weaker than those observed at intermediate levels of sequence identity. One possible reason for the weaker differences in this range is the greater uncertainty in sequence and structure alignments at less than 30% sequence identity,[37,39,49,50] which can lead to less accurate predictions of orthology and paralogy, as well as less accurate assessments of sequence similarity and structural divergence.

As expected, little difference was found between orthologs and paralogs above 70% sequence identity (average C_α RMSDs of 0.77 and 0.72 Å, respectively). We can reliably generate high-accuracy alignments for proteins with high sequence identities. In contrast, sequence-based detection of orthology and paralogy differentiates between the two on the basis of differences in sequence similarity, thus becoming more difficult for groups of very similar sequences. In addition, underlying our central hypothesis is the expectation that in the presence of fewer functional constraints, a protein will be free to acquire more sequence changes that have a marked effect on its structure. Evidence of this effect will be less apparent from proteins that are very similar in sequence.

As alignment methods and methods for orthology and paralogy assignment improve, additional studies could further test our hypothesis. Improvements in computing power alone should allow larger-scale application of tree-based methods for assigning

orthology and paralogy. In addition, as new genomes continue to be sequenced, additional data will become available for a more comprehensive analysis.

Abstraction of general principles from individual cases

We tried to determine, by inspecting many cases, common structural features that lead to greater structural similarity among orthologs compared to paralogs. Although the overall trends in the data are statistically significant, there were also many exceptions [e.g., Figs. 4(b) and 5(b)]. Therefore, we also inspected cases in which paralogs had greater structural similarity than orthologs. However, despite being able to rationalize to some degree the observed sequence and structure differences in individual cases, we were not able to discern any general principles that would allow us to predict when individual cases would conform to our hypothesis. In fact, there is no guarantee that there are such general principles, other than the laws of physics that determine how a protein sequence folds to its native structure.

Significance of results

The small confidence intervals shown in Figures 1 and 2 are in part due to the large sample sizes available for our analyses. As described later, these confidence intervals were calculated based on the Student *t*-distribution, allowing us to use the sample standard deviation to estimate the intervals. These intervals do not assume a particular underlying distribution of the RMSDs between orthologous or paralogous pairs. The calculation of the confidence intervals accounts for the variance in the samples, and thereby also accounts for any nonsystematic errors in the determination of crystallographic structures, the alignment process, or the determination of orthology. Some protein families are more or less abundant than others in SCOP and therefore in our data set. Although this uneven distribution of protein families certainly affects the average RMSDs determined in our analysis, in the absence of a clear framework for determining how protein space should be sampled, we used all available pairs of proteins. Deviations from the general trends shown by the curves in Figures 1 and 2 may be due in part to this uneven representation of protein families at different sequence identities, to difficulty in correctly classifying orthologs and paralogs at high sequence identities, or to smaller differences between orthologs and paralogs at high sequence identities (discussed earlier).

We recognize a difference between statistical significance of a difference between two samples (which depends on the sizes of the samples) and practical utility of the difference for predictive purposes (which depends on its magnitude). We suggest that it is in the middle range of sequence identities (30–70%) that using evolutionary relationships is most useful as an adjunct to using sequence identity to estimate structural similarity; in this range, the difference appears large enough to have practical utility, for example in the selection of templates for comparative modeling, as discussed next.

Implications for comparative modeling

We have shown that for a large set of orthologs (86,676 pairs of domains) and paralogs (197,783 pairs of domains), orthologs sharing sequence identities below 70% are more structurally similar than paralogs at a similar level of sequence identity (see Fig. 1). These results have implications for comparative modeling. The accuracy of any comparative model is directly dependent on the structural similarity between the target and the template. Because sequence similarity is frequently used as a predictor of structural similarity, the protein with the highest sequence similarity to the target is often chosen as the template. Our results show that combining knowledge of orthology or paralogy with sequence similarity provides a better predictor of structural similarity, and thus, of the best template for modeling. In the range of target-template sequence identities below 70%, using an ortholog is therefore expected to give better results on average than using a paralog. Better results are expected even when the sequence identity between the target and its ortholog is lower than the sequence identity between the target and a paralog by up to 17 percentage points. We suggest that for sequence identities below 70%, the choice of templates for comparative modeling should be based not only on sequence identity, but also take into account the evolutionary relationship between the target and possible templates.

Predicting protein function is a difficult task, and when looking to comparative models for clues about function, accuracy of the modeled functional residues becomes critical. We found that functional residues in pairs of orthologous proteins were more structurally similar (up to 0.26 Å lower in average RMSD) than functional residues in pairs of paralogs in the same range of sequence similarity (see Fig. 2). Therefore, using orthologous templates becomes even more important when accurate modeling of functional residues is critical, such as when using models to predict function or for computational docking of ligands.[51–53]

Implications for structural genomics projects

Our results also have implications for target selection for structural genomics projects, and more generally when attempting to determine which protein structures would provide the most complete coverage of protein structure space.[37,44,54–56] High-priority proteins for structural determination are often identified as those having low sequence similarity to any previously solved protein structure. However, our results show that predictions of whether a pair of proteins is orthologous or paralogous can significantly change the expected structural similarity between the two. Thus,

known orthologous or paralogous relationships between candidate proteins for structure determination and known structures should ideally be factored in when prioritizing structures to be solved.

Future directions

Here, we addressed the question of whether or not sequence-structure relationships that had been found to apply to general classes of proteins were quantitatively different for orthologs versus paralogs. We can make our study even more specific by dividing the set of examined proteins not only by orthology or paralogy, but also by additional attributes such as fold class, superfamily, or family; by length; or by any other physicochemical properties. Additionally, as our hypothesis was based on the idea that it is functional similarity between orthologs that leads to their more similar structures at a given level of sequence identity, separating by functional attributes those groups of proteins that are to be analyzed makes sense.

Materials and Methods

We required sequence identities and structural similarities between pairs of domains of known evolutionary relationship (orthologous or paralogous). Next, we describe our methods for obtaining data sets of whole domains with known structures, identifying evolutionary relationships (orthology or paralogy) among these domains, identifying the ligand-binding residues of the domains, aligning sequences to calculate sequence identity, and obtaining the structural superpositions necessary to calculate structural similarity.

Protein domains of known structure

We focused our analyses on single domains to avoid the difficulties inherent in large-scale, automated comparisons of multi-domain structures; these difficulties arise from differences in the relative positions and orientations of their domains. Our data set consisted of all domains in the manually curated SCOP 1.69 database of protein domains[57] for which the full protein sequence was identical to all or part of a gene sequence listed in the OrthoMCL DB V1 database of ortholog group predictions for 55 complete genomes.[36] If multiple SCOP domains matched a single sequence in OrthoMCL DB, a single representative SCOP domain was chosen. Whenever possible, the representative domain was from the same species as the OrthoMCL DB gene sequence. Otherwise, the highest-resolution crystallographic structure from any species was chosen. When no crystallographic structure was available, an NMR structure was used. Although data sets filtered by different criteria might yield different results, we chose this one to have as large a sample as reasonably possible. Restricting the data set to crystallographically determined structures with resolutions better than 2.5 Å and to those with resolutions better than 2.0 Å gave very similar results to those presented earlier.

Evolutionary relationships between pairs of protein domains

We adopted orthology and paralogy definitions from OrthoMCL DB, which uses whole-genome alignments to determine clusters of orthologous groups. The OrthoMCL method[28,58] overcomes the inability of simple reciprocal best hit approaches to detect many-to-many relationships by including bridging in-paralogous relationships (arising from duplication events subsequent to species divergence). Ortholgous relationships are detected using comparisons of full-length protein sequences. We labeled pairs of domains as orthologous when both domains in the pair were from different species and belonged to the same set of orthologous groups in OrthoMCL DB. Pairs of domains were labeled as paralogous whenever they were from the same species and not in the same OrthoMCL DB groups. This process resulted in 86,676 pairs of orthologous domains and 197,783 pairs of paralogous domains. Among the orthologous pairs, 8,277 distinct SCOP domains were included, and among the paralogous pairs, 8,765 distinct SCOP domains were included (of 70,859 available SCOP domains).

Ligand-binding residues

For each domain, we used the annotations in LigBase,[47] a structural database of aligned ligand binding sites, to determine which residues bound ligands. Ligand-binding residues are defined as those residues with at least one atom within 5 Å of any ligand atom in an experimentally determined structure. These ligands include small molecules, such as metal ions, nucleotides, and peptides, but exclude nucleic acids and other proteins.

Sequence alignments

Coordinate files and sequences for the studied protein domains were taken from the ASTRAL compendium for sequence and structure analysis.[59] We used sequence identity as a measure of the similarity between pairs of sequences because sequence identity is a commonly used and well understood measure that correlates well with structural similarity above 30% sequence identity.[1,6] Other measures of sequence similarity were also used, such as sequence similarity calculated using the BLOSUM 62 substitution matrix,[60,61] but did not change our conclusions (data not shown). To calculate sequence identities between pairs of domains, structure-based pairwise alignments were constructed using three different methods: *align3d*, available as part of Modeller 9v2,[62] CE,[63] and TM-Align.[64] For each pair of domains, we selected the best resulting alignment, defined as the alignment with the greatest number of equivalent positions (i.e., the

number of aligned residue pairs with C_α atoms within 3.5 Å of each other when the domain structures were superposed). When multiple programs produced alignments that were equivalent by this measure, the alignment with the lowest pairwise RMSD upon structural superposition was chosen. When all three alignments had the same RMSD, we selected the alignment obtained using *align3d*. These alignments of domain pairs were used both for superposing whole domains and for superposing sets of ligand-binding residues.

Structure superpositions

We used Modeller's superpose method to create pairwise structural superpositions. C_α RMSDs between pairs of aligned protein domains were calculated using all equivalent positions (as defined earlier), and the resulting superposition for any pair of domains was the one that minimized this C_α RMSD. Superpositions of aligned ligand-binding residues were calculated to minimize RMSD over all atoms in those ligand-binding residues only (i.e., the remaining residues in those domains did not affect the superposition). The residues superposed included those from all alignment positions in which the residue from the query domain was determined to be a ligand-binding residue. RMSDs for ligand-binding residues were not included in the analysis if there were fewer than three such alignment positions.

Statistical analysis

Two-sided 95% confidence intervals for the mean RMSDs were calculated using the Student *t* distribution,[65] using sample standard deviations to estimate the intervals.

Acknowledgments

The authors are grateful to Ranyee Chiang, Eswar Narayanan, and Sunil Ojha for discussion about this project. They acknowledge the support from Mike Homer, Ron Conway and hardware gifts from IBM, Intel, Hewlett-Packard, and NetApp.

References

1. Chothia C, Lesk AM (1986) The relation between the divergence of sequence and structure in proteins. EMBO J 5:823–826.
2. Hubbard TJ, Blundell TL (1987) Comparison of solvent-inaccessible cores of homologous proteins: definitions useful for protein modelling. Protein Eng 1:159–171.
3. Flores TP, Orengo CA, Moss DS, Thornton JM (1993) Comparison of conformational characteristics in structurally similar protein pairs. Protein Sci 2:1811–1826.
4. Russell RB, Saqi MA, Sayle RA, Bates PA, Sternberg MJ (1997) Recognition of analogous and homologous protein folds: analysis of sequence and structure conservation. J Mol Biol 269:423–439.
5. Sauder JM, Arthur JW, Dunbrack RL, Jr (2000) Large-scale comparison of protein sequence alignment algorithms with structure alignments. Proteins 40:6–22.
6. Wilson CA, Kreychman J, Gerstein M (2000) Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores. J Mol Biol 297:233–249.
7. Rost B (1999) Twilight zone of protein sequence alignments. Protein Eng 12:85–94.
8. Yang AS, Honig B (2000) An integrated approach to the analysis and modeling of protein sequences and structures. II. On the relationship between sequence and structural similarity for proteins that are not obviously related in sequence. J Mol Biol 301:679–689.
9. Babbitt PC (2003) Definitions of enzyme function for the structural genomics era. Curr Opin Chem Biol 7: 230–237.
10. Tipton K, Boyce S (2000) History of the enzyme nomenclature system. Bioinformatics 16:34–40.
11. Devos D, Valencia A (2000) Practical limits of function prediction. Proteins 41:98–107.
12. Rost B (2002) Enzyme function less conserved than anticipated. J Mol Biol 318:595–608.
13. Tian W, Skolnick J (2003) How well is enzyme function conserved as a function of pairwise sequence identity? J Mol Biol 333:863–882.
14. Joshi T, Xu D (2007) Quantitative assessment of relationship between sequence similarity and function similarity. BMC Genomics 8:222.
15. Sangar V, Blankenberg DJ, Altman N, Lesk AM (2007) Quantitative sequence-function relationships in proteins based on gene ontology. BMC Bioinformatics 8:294.
16. Hegyi H, Gerstein M (1999) The relationship between protein structure and function: a comprehensive survey with application to the yeast genome. J Mol Biol 288:147–164.
17. Todd AE, Orengo CA, Thornton JM (2001) Evolution of function in protein superfamilies, from a structural perspective. J Mol Biol 307:1113–1143.
18. Thornton JM, Orengo CA, Todd AE, Pearl FM (1999) Protein folds, functions and evolution. J Mol Biol 293: 333–342.
19. Whisstock JC, Lesk AM (2003) Prediction of protein function from protein sequence and structure. Q Rev Biophys 36:307–340.
20. Shakhnovich BE, Harvey JM (2004) Quantifying structure-function uncertainty: a graph theoretical exploration into the origins and limitations of protein annotation. J Mol Biol 337:933–949.
21. Koehl P, Levitt M (2002) Sequence variations within protein families are linearly related to structural variations. J Mol Biol 323:551–562.
22. Wood TC, Pearson WR (1999) Evolution of protein sequences and structures. J Mol Biol 291:977–995.
23. Fitch WM (1970) Distinguishing homologous from analogous proteins. Syst Zool 19:99–113.
24. Sonnhammer EL, Koonin EV (2002) Orthology, paralogy and proposed classification for paralog subtypes. Trends Genet 18:619–620.
25. Koonin EV (2005) Orthologs, paralogs, and evolutionary genomics. Annu Rev Genet 39:309–338.
26. Tatusov RL, Koonin EV, Lipman DJ (1997) A genomic perspective on protein families. Science 278:631–637.
27. Hulsen T, Huynen MA, de Vlieg J, Groenen PM (2006) Benchmarking ortholog identification methods using functional genomics data. Genome Biol 7:R31.
28. Chen F, Mackey AJ, Vermunt JK, Roos DS (2007) Assessing performance of orthology detection strategies applied to eukaryotic genomes. PLo S ONE 2:e383.
29. Mirkin B, Muchnik I, Smith TF (1995) A biologically consistent model for comparing molecular phylogenies. J Comput Biol 2:493–507.

30. Page RDM, Charleston MA (1997) From gene to organismal phylogeny: reconciled trees and the gene tree/species tree problem. Mol Phylogenet Evol 7:231–240.
31. Zhang L (1997) On a Mirkin-Muchnik-Smith conjecture for comparing molecular phylogenies. J Comput Biol 2: 177–187.
32. Eulenstein O, Mirkin B, Vingron M (1998) Duplication-based measures of difference between gene and species trees. J Comput Biol 5:135–148.
33. Li H, Coghlan A, Ruan J, Coin LJ, Heriche JK, Osmotherly L, Li R, Liu T, Zhang Z, Bolund L, et al. (2006) TreeFam: a curated database of phylogenetic trees of animal gene families. Nucleic Acids Res 34:D572–D580.
34. Koonin EV, Fedorova ND, Jackson JD, Jacobs AR, Krylov DM, Makarova KS, Mazumder R, Mekhedov SL, Nikolskaya AN, Rao BS, et al (2004) A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. Genome Biol 5:R7.
35. Tatusov RL, Federova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN, et al. (2003) The COG database: an updated version includes eukaryotes. BMC Bioinformatics 4:41
36. Chen F, Mackey AJ, Stoeckert CJ, Jr, Roos DS (2006) OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. Nucleic Acids Res 34: D363–D368.
37. Baker D, Sali A (2001) Protein structure prediction and structural genomics. Science 294:93–96.
38. Petrey D, Honig B (2005) Protein structure prediction: inroads to biology. Mol Cell 20:811–819.
39. Ginalski K (2006) Comparative modeling for protein structure prediction. Curr Opin Struct Biol 16:172–177.
40. Misura KM, Chivian D, Rohl CA, Kim DE, Baker D (2006) Physically realistic homology models built with ROSETTA can be more accurate than their templates. Proc Natl Acad Sci USA 103:5361–5366.
41. Eswar N, Eramian D, Webb B, Shen MY, Sali A (2008) Protein structure modeling with MODELLER. Methods Mol Biol 426:145–159.
42. Kim SH (1998) Shining a light on structural genomics. Nat Struct Biol 5 (Suppl):643–645.
43. Sanchez R, Pieper U, Melo F, Eswar N, Marti-Renom MA, Madhusudhan MS, Mirkovic N, Sali A (2000) Protein structure modeling for structural genomics. Nat Struct Biol 7 (Suppl):986–990.
44. Bray JE, Marsden RL, Rison SC, Savchenko A, Edwards AM, Thornton JM, Orengo CA (2004) A practical and robust sequence search strategy for structural genomics target selection. Bioinformatics 20:2288–2295.
45. Todd AE, Marsden RL, Thornton JM, Orengo CA (2005) Progress of structural genomics initiatives: an analysis of solved target structures. J Mol Biol 348:1235–1260.
46. Chandonia JM, Brenner SE (2006) The impact of structural genomics: expectations and outcomes. Science 311: 347–351.
47. Stuart AC, Ilyin VA, Sali A (2002) LigBase: a database of families of aligned ligand binding sites in known protein sequences and structures. Bioinformatics 18:200–201.
48. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE (2004) UCSF Chimera–a visualization system for exploratory research and analysis. J Comput Chem 25:1605–1612.
49. Sanchez R, Sali A (1998) Large-scale protein structure modeling of the *Saccharomyces cerevisiae* genome. Proc Natl Acad Sci USA 95:13597–13602.
50. Kryshtafovych A, Venclovas C, Fidelis K, Moult J (2005) Progress over the first decade of CASP experiments. Proteins 61 (Suppl 7):225–236.
51. Shoichet BK (2004) Virtual screening of chemical libraries. Nature 432:862–865.
52. Huang N, Kalyanaraman C, Bernacki K, Jacobson MP (2006) Molecular mechanics methods for predicting protein-ligand binding. Phys Chem Chem Phys 8:5166–5177.
53. Huang N, Jacobson MP (2007) Physics-based methods for studying protein-ligand interactions. Curr Opin Drug Discov Dev 10:325–331.
54. Chandonia JM, Brenner S (2005) Update on the pfam5000 strategy for selection of structural genomics targets. Conf Proc IEEE Eng Med Biol Soc 1:751–755.
55. Chandonia JM, Kim SH, Brenner SE (2006) Target selection and deselection at the Berkeley Structural Genomics Center. Proteins 62:356–370.
56. Minary P, Levitt M (2008) Probing protein fold space with a simplified model. J Mol Biol 375:920–933.
57. Murzin AG, Brenner SE, Hubbard T, Chothia C (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 247:536–540.
58. Li L, Stoeckert CJ, Jr, Roos DS (2003). OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res 13:2178–2189.
59. Chandonia JM, Hon G, Walker NS, Lo Conte L, Koehl P, Levitt M, Brenner SE (2004) The ASTRAL compendium in 2004. Nucleic Acids Res 32:D189–D192.
60. Henikoff S, Henikoff JG (1992) Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci USA 89:10915–10919.
61. Styczynski MP, Jensen KL, Rigoutsos I, Stephanopoulos G (2008) BLOSUM62 miscalculations improve search performance. Nat Biotech 26:274–275.
62. Sali A, Blundell TL (1993) Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol 234:779–815.
63. Shindyalov IN, Bourne PE (1998) Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. Protein Eng 11:739–747.
64. Zhang Y, Skolnick J (2005) TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res 33:2302–2309.
65. Hogg RV, Ledolter J (1987) Engineering Statistics. New York: Macmillan Publishing Company, p 420.