Evolutionary Rate at the Protein Domain Level is Constrained by Importance to Network Dynamics

EVOLUTIONARY RATE AT THE PROTEIN DOMAIN LEVEL IS CONSTRAINED BY IMPORTANCE TO NETWORK DYNAMICS

Draft produced May 1, 2012

By

BRIAN KENDALL MANNAKEE

A Thesis Submitted to The Honors College

In Partial Fulfillment of the Bachelors degree

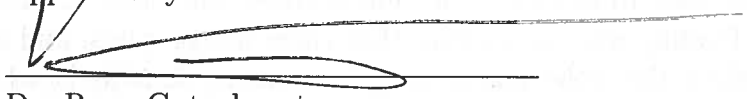
With Honors in

Biochemistry and Molecular Biophysics

THE UNIVERSITY OF ARIZONA

MAY 2012

Approved by:



Dr. Ryan Gutenkunst
Department of Molecular and Cellular Biology

The University of Arizona Electronic Theses and Dissertations
Reproduction and Distribution Rights Form

Name (Last, First, Middle) Brian K. Mannakee	
Degree title (eg BA, BS, BSE, BSB, BFA): BS	
Honors area (eg Molecular and Cellular Biology, English, Studio Art): Biochemistry	
Date thesis submitted to Honors College:	
Title of Honors thesis: Evolutionary Rate at the protein Domain Level is constrained by importance to network dynamics	
:The University of Arizona Library Release	I hereby grant to the University of Arizona Library the nonexclusive worldwide right to reproduce and distribute my dissertation or thesis and abstract (herein, the "licensed materials"), in whole or in part, in any and all media of distribution and in any format in existence now or developed in the future. I represent and warrant to the University of Arizona that the licensed materials are my original work, that I am the sole owner of all rights in and to the licensed materials, and that none of the licensed materials infringe or violate the rights of others. I further represent that I have obtained all necessary rights to permit the University of Arizona Library to reproduce and distribute any nonpublic third party software necessary to access, display, run or print my dissertation or thesis. I acknowledge that University of Arizona Library may elect not to distribute my dissertation or thesis in digital format if, in its reasonable judgment, it believes all such rights have not been secured. Signed: _____ Date: 5-2-2012

Last updated: Nov 15, 2009

Abstract

A fundamental question for evolutionary biology is why different proteins evolve at dramatically different rates. As the evolutionary time between two species and their common ancestor increases they accumulate a proportional number of amino acid changes between homologous proteins, but this proportional increase is not the same for all proteins. This difference in rate is attributed to the action of natural selection. While intuition suggests that natural selection acts most strongly to preserve and improve the function of proteins, little evidence supports this idea. Instead, the strongest predictor of protein evolutionary rate found to date is protein expression level, suggesting that selection acts primarily to prevent protein mis-folding. Here we suggest this apparent contradiction arises because the methods used to measure protein functional importance are ill-suited to capture the subtle interplay between protein structure and function. We introduce a measure of functional importance called Dynamical Influence, leveraging computational models of cellular systems for a much finer view of the functional importance of a protein in its network context. Comparing this measure with protein evolutionary rates across the vertebrate evolutionary tree we find strong evidence that selection operates to preserve and refine protein function to a much greater extent than previously observed.

Introduction

Background

A principal problem in evolutionary biology is to understand the mechanism by which proteins evolve and the reason that they evolve at different rates [20]. In 1962 Emil Zuckerkandl and Linus Pauling observed that amino acid substitutions in hemoglobin across different lineages appeared to accumulate linearly in time. In other words, as the time separating two species from their most recent common ancestor increases, the number of amino acid differences in their hemoglobin increases proportionally [25]. This observation led them to hypothesize that there is a molecular clock, or constant rate at which proteins acquire slightly advantageous mutations, and that this clock can be used to estimate the time elapsed since two species diverged. Zuckerkandl and Pauling were also aware that there are proteins, and even portions of proteins, for which the molecular clock either runs much more slowly, or not at all. For instance, at the time they wrote 11 amino acids(approx 10%) were known to be strictly invariant for all species for which globins had been sequenced, and they proposed that the effect of natural selection was suppressing the rate of

evolution at these amino acids.

In 1968 Motoo Kimura proposed the neutral theory of evolution which provides a theoretical framework for understanding the molecular clock as well as the apparent differences in the rate of evolution that arise within proteins and across different lineages [13]. Because the mechanism of DNA replication is prone to some small level of error, mutations repeatedly arise in the genome, and the rate at which these errors arise is essentially constant. The neutral theory says that most of these mutations are neutral, meaning that they have no effect on the organism. Since they arise in a single organism and provide no selective advantage, most of these mutations do not spread through the population and are quickly lost. However, Kimura was able to show statistically that some small portion of these neutral mutations will persist and eventually rise to become the dominant allele in the population. This process of neutral evolution is stochastic in nature and the rate at which neutral mutations become fixed in a population depends only on the rate at which random mutations occur in the genome due to replication errors. The neutral theory of molecular evolution has been tested extensively and makes highly accurate predictions about the speed of the molecular clock. These predictions allow us to identify proteins for which the molecular clock is running at a pace not consistent with neutral evolution and to measure by how much the rate differs from neutral. The theory suggests that natural selection must be acting on these proteins, and the slower a protein is evolving the more selective constraint it is under.

Mutations in the genome occur when one nucleotide in the DNA is substituted with another, for instance a G becomes a T. Because multiple codons can code for the same amino acid, these nucleotide substitutions will have one of two outcomes. First, the resulting codon may encode the same amino acid and the protein translated from the gene is unchanged. This is a synonymous mutation. The second possible result is that the new codon may encode for a different amino acid and the protein encoded by the gene is changed. This second type of mutation is called a non-synonymous mutation. Synonymous mutations are largely neutral, and thus occur at a rate given by the neutral theory. Non-synonymous mutations are typically not completely neutral, and thus rise to fixation in a population at a rate that depends on the selective advantage or disadvantage they provide the organism. The ratio of the rates at which non-synonymous(dN) and synonymous(dS) substitutions occur in a gene gives a measure of the selective pressure the gene is under. For a gene under purifying selection, synonymous substitutions will occur at a higher rate than non-synonymous because natural selection is acting to suppress changes in protein sequence, and the ratio dN/dS will be less than one. In the present work we will use dN/dS as a measure of evolutionary rate in order to capture the magnitude of

the selective pressure on the proteins we are studying. Identifying the mechanism or mechanisms by which natural selection is acting on these proteins has been the subject of intense study for decades. To date several mechanisms have been proposed, but none explains all of the observed difference in evolutionary rate. [18]

Three categories of variables correlate with evolutionary rate

The first category, and the one with the strongest correlation to evolutionary rate, are those variables that involve the rate at which genes are translated in the cell. Drummond et.al. showed that in the yeast S. cerevisiae expression level, measured as mRNA abundance in a cell, correlates strongly with their evolutionary rate. Proteins that are more highly expressed evolve more slowly [3]. By controlling for several confounding factors they showed that expression level alone explains roughly half of the variation in evolutionary rate. They have shown experimentally that as the rate of expression increases selection for robustness to translational missense errors increases, because protein misfolding places a heavy burden on the cell [7].

The second category of variable that shows a correlation with evolutionary rate are those dealing with protein network structure. These include the number of interaction partners a protein has, and its centrality within the protein interaction network. Early work on the effect of network function looked at the number of interaction partners a protein had and found a very weak correlation between number of interaction partners and evolutionary rate [6]. This correlation was subsequently found to be statistically insignificant, and more recent work has focused on the centrality of a protein in its cellular network [23]. Centrality is measured by first determining the topology of a protein interaction network, where each protein represents a node in the network. It is then possible to calculate the shortest distance between each pair of proteins in the network. The centrality of a protein is determined by counting the number of times a protein lies along the shortest path between two other proteins. The more interaction pathways a protein lies along, the more central it is to the network, and the higher centrality score it is assigned. Hahn and Kern have shown that, across the yeast, nematode, and fruit fly genomes, proteins that are more central to their networks evolve more slowly [11].

The final category of variable that correlates with evolutionary rate measures the function of a protein. Function would logically be expected to correlate strongly with evolutionary rate, as it seems clear that a protein whose function is extremely important for the organism would have little room for change without having a deleterious effect, and this should cause functionally important proteins to evolve more slowly. Functional importance is typically measured as protein dispensability, or the

survival and growth rates achieved by an organism when the gene for a protein has been completely removed from the genome. Hirsh et al. estimated dispensability by measuring the fitness effect of single gene deletions in S. cerevesiae and found that genes that were less dispensable, i.e. whose deletion had a larger fitness effect, evolved more slowly [12]. Surprisingly, the correlation they found between function and evolution was fairly weak, and subsequent work by Drummond et al. [4] has shown rather convincingly that when expression level is controlled for, the correlations between evolutionary rate and all network and functional variables become negligible. Expression appears to be the dominant determinant of evolutionary rate in yeast.

The picture in mammals and vertebrates is somewhat more complex as evolutionary experiments are more difficult to perform and genetic data for a wide range of species and proteins is only slowly becoming available. To date vertebrate evolutionary rates have been shown to correlate best with tissue specificity [16], systemic factors such as gene age and complexity [22], and translation rates [5]. None of these correlations are as strong as those in yeast, and in no case has function been found to correlate strongly with evolutionary rate when controlled for other factors. So while theory and intuition predict that a protein's functional importance would apply significant evolutionary constraint on its rate of evolution, evidence to date suggests precisely the opposite. Here we suggest that the cause of this discrepancy lies in the choice of variables used to measure functional importance. We propose a new measure of functional importance, called Dynamical Influence, which provides a much finer measure of the importance of a protein to the function of an organism.

Dynamical Influence

The dynamical influence of a protein is a measure of the functional importance of that protein within its cellular network context. There is a large systems biology literature in which researchers combine precise biochemical rate constant information with detailed knowledge about the interactions of different proteins to build large computational models of cellular networks. We leverage these sophisticated computational models by using integration techniques to measure the sensitivity of the network to changes in the reaction rates for the reactions in which each protein participates. Dynamical influence quantifies the magnitude of the effect exerted on the network by changes in a protein's reaction constants.

Relations between protein structure, function, and evolutionary rate

In previous studies attempting to link function with evolutionary rate the evolutionary rate was measured for the entire protein. But individual domains within a protein, which have different functions and participate in separate reactions, evolve at distinctly different rates. For example, Sos1 is an activator of the Ras protein which plays a major role in various cell signaling networks, and the cdc25 domain of Sos1 is where Ras activation occurs. The Grb2 binding domain is the domain which brings the protein into position to interact with Ras. The entire protein has an evolutionary rate, measured as the ratio of non-synonymous to synonymous amino acid substitutions over the vertebrate tree, of 0.16. However, when we measure the evolutionary rate for these two functionally important domains, we find the cdc25 domain has an evolutionary rate of .01 and the Grb2 binding domain has an evolutionary rate of .03, each much slower than for the entire protein. This is possible because the linker regions between these domains are relatively less constrained functionally and structurally, and so are able to evolve more rapidly. Thus, when the evolutionary rate is measured across the entire protein substantial information is lost. If evolutionary pressure related to a specific activity for a protein would be directed at the domain at which the activity occurs, it is important to measure the evolutionary rate for that specific domain when looking for signatures of evolutionary constraint based on function.

Constraints on protein evolution imposed by functional requirements

Protein evolution proceeds through random mutations in the genome that create changes in the amino acid sequences of proteins [13]. Protein structure is controlled by amino acid sequence and function is intimately tied to structure at all levels. At the level of primary structure, or amino acid sequence, the effect of a single amino acid change can be profound. For instance, changing a single catalytic amino acid in the active site of an enzyme may completely destroy its function. Changing a single amino acid at the site where the protein binds another protein may completely disrupt their association with disastrous consequences for the cell. The secondary structure of proteins is controlled by precise local interactions between amino acids, and changing a single amino acid can alter local structure in complex ways affecting the function of the protein at crucial catalytic and binding sites. Finally, the tertiary structure of a protein is the result of interactions between amino acids that may be far apart in its' amino acid sequence.

Changes in a small number of amino acids can have large effects on the overall structure of the protein, causing catalytic sites to become blocked, or binding sites

to shift away from their proper position. They can also have smaller and more subtle effects like slightly changing the position of a catalytic amino acid in an enzyme so that the reaction, while is still occurs, occurs at a dramatically different rate than that in the protein prior to the mutation. Gureasko et.al. have done single amino acid mutation studies in Sos1, measuring the effect on Ras activation of a large number of mutations. They find that Ras activation is exquisitely sensitive to substitutions even when they are distant in the sequence from the active site. Many substitutions make small changes to the activation rate, both increasing and decreasing it without destroying the activity [8]. These small changes in activity may improve the function of a protein or degrade it, but it seems logical that the more important the activity of a protein domain is in the context of its network structure, the less it will be able to tolerate even small changes in catalytic or binding activity. Thus an evolutionary constraint should be applied to functionally important proteins such that they evolve more slowly than less important proteins. Here we evaluate ten models of protein interaction networks and find strong evidence that the more dynamical influence a protein domain exerts in its cellular network context, the slower that domain evolves. We also show that this effect appears to be independent of other previously noted correlations with expression level and network structure.

Methods

Calculation of Dynamical Influence

Detailed mammalian biochemical models are obtained from the Biomodels database [14]. Each model is loaded into a Python package called SloppyCell [9] [19] and the dynamical influence of each reaction parameter in the model is calculated as follows:

Consider the sample protein interaction network model in Figure 1. Each parameter k is the rate constant for a particular reaction in the model, and for each species y in the model we can calculate concentration as a function of time and the value of any parameter k_i

$$\frac{dy}{dt} = f(y, t; k_i) \tag{1}$$

The dynamical influence κ_i of parameter k_i is a measure of the effect from a small change in that parameter on each species in the model, summed over every species

$$\kappa_i^2 \propto \sum_y \frac{1}{\sigma_y^2} \int \left(\frac{\partial y_c(t,k)}{\partial log k_i} \right)^2 \Big|_{k=k_0} dt \tag{2}$$

The dynamical influence of a protein or protein domain is then the geometric average of the κ's for all of the reactions in which that protein or domain participates.

Calculation of evolutionary rate

Protein evolutionary rates were calculated across the vertebrate evolutionary tree in Figure 2. Homologous protein alignments were downloaded from the NCBI Homologene database, and only proteins having sequences for at least six of the species in Figure 2 were included in the analysis. All protein evolutionary rates were calculated using the codeml package written by Yao et al [24], which gives dN/dS ratios for protein sequence alignments along a given tree. Whole protein evolutionary rates were calculated using the codeml setting Mgene=0 and rates for individual protein domains were calculated using codeml setting Mgene=3.

Assignment of reaction constants to proteins and protein domains

The proteins involved in each reaction were obtained from the biochemical models, using the NCBI Uniprot identifiers provided by the modellers. Extensive biochemical literature review was conducted in order to assign reactions to specific domains in each protein. Where proteins have a single domain, or the exact reaction site was unknown, the reaction was assigned to the entire protein. Each reaction was also identified as either catalytic or non-catalytic so that these two reaction types could be analyzed separately. We created a Python package called DynEvol which serves as the bioinformatics infrastructure of the project. The package maintains all of the relevant data including protein sequence information, expression and network data, biochemical model specifications, dynamical influence data, and reaction/domain assignments [17].

Statistical Analysis

All statistical analysis and plotting were done using Python. Correlation coefficients ρ were calculated using the spearman rank correlation algorithm, which calculates the correlation between the ranks of two variables and is less sensitive to outliers in the tail of the data. Principal component analysis was done using the R princomp() function accessed from Python via the Rpy2 interface.

Expression, network structure, and essentiality data

Mouse gene expression data was obtained from the GNF1M mouse database [21]. Expression and specificity were calculated using the procedure from Drummond et. al. in Cell(2008) [5]. Mouse phenotype data were obtained from the Mouse Genome Informatics MGD database [1], and essentiality was calculated from phenotypes as described in Liao et. al. (MBE 2006) [15]. Protein interaction data were downloaded from the Interologous Interaction Database maintained by the Jurisica lab at University of Toronto [2], and protein degree and betweenness-centrality were calculated using the Python package NetworkX [10]

Results

Correlation and partial correlation at the protein domain level

Figure 3 and Table 1 show the results of correlation analysis between evolutionary rate, dynamical influence, and expression level. In 8 of 10 models evaluated we find a negative correlation between evolutionary rate and dynamical influence at the domain level, and this correlation remains after controlling for the expression level of each protein.

Table 2 shows the correlations found for just those domains participating in catalytic reactions. Since catalytic reactions often occur at evolutionarily highly conserved domains whose function is highly sensitive to small changes in amino acid sequence we would expect these domains in general to have very low evolutionary rates, and for those rates to be more strongly correlated with expression levels than function. The data in Table 2 supports this intuition. The correlations between dynamical influence and evolutionary rate when controlled for expression are generally lower and less consistent than the same correlations for the combined reactions.

Non-catalytic domains on the other hand mediate protein interactions, and are typically less sensitive to small changes in amino acid sequence, making them more likely to accumulate small changes that effect the dynamics of the cellular network. Table 3 shows the correlations for domains involved in non-catalytic reactions. We would expect to see higher evolutionary rates in general for these domains, and more correlation with dynamic influence than expression, and we find that to be the case.

Principal component analysis

As Drummond et al. point out [5], in the presence of noise partial correlation analysis can lead to spurious results, showing correlations where none actually exist. They suggest principal component analysis as a method to circumvent the potential problems of partial correlation analysis. From each of the three categories of variable that correlate with evolutionary rate we chose the variable with the most consistent proven correlation, and performed a principal component analysis with those variable. Table 4 shows the result of our analysis. We find that the component that explains about half the variance in the data is comprised solely of network variables. Much of the rest of the variance is explained by two components, one primarily consisting of expression and the other of dynamical influence.

Discussion

Past work by Drummond et.al. [5] has shown a clear effect on evolutionary rate from expression levels due to translation-misfolding costs, across a large variety of species. They have also cast doubt on the effect of protein function on evolutionary rate. The work of Hirsh et.al making use of protein dispensability/lethality data as a proxy for protein functional importance [12] appears to show a weak correlation between dispensability and evolutionary rate, but this disappears when controlled for expression. Similarly, work by Fraser et.al looking at protein interaction network structure as a proxy for protein functional importance [6], which found a weak correlation between the number of interaction partners and evolutionary rate also disappears when controlled for expression level. These failures led to the somewhat counter-intuitive hypothesis that translation-induced misfolding may be the dominant factor in determining evolutionary rate, and that function may play either a much smaller role or no role at all. While many in the evolutionary biology community have been puzzled by the inability to find a role for function in protein evolution, the strength of the evidence for expression level as the sole determinant of evolutionary rate differences has appeared overwhelming.

Here we present evidence that, at least in vertebrates, function may have a substantially larger effect than has been found previously. When the detailed biochemical context in which proteins carry out their function is taken into account, the effect of protein function on evolution becomes clearer. Calculating protein functional importance by looking at the effect of changes in the function of the protein on the cellular network in which it operates, we get a crisper picture of that proteins importance than we get from either knockout lethality or counting interaction part-

ners. Further sharpening our analysis by calculating both influence and evolutionary rate at the domain level, we focus on the parts of the protein at which evolutionary selection may be operating. When measured this way, we find a strong correlation between protein function and evolutionary rate, and this correlation remains strong when controlled for expression level. Principal component analysis suggests that the magnitude of the two effects are similar. Additionally, the results pass a simple test of logical consistency by behaving as the biochemistry would suggest when breaking protein domains down into the categories of catalytic and non-catalytic. By taking a fresh approach to thinking about how to measure the functional importance of a protein, this work provides a catalyst for the field to re-evaluate many of the strong results showing that function plays little or no role in protein evolution.

Future directions

A substantially larger number of detailed biochemical models is needed to verify the effects we have shown here. More models will help increase our understanding of the sources of differences in evolutionary rate among proteins. We hope that as the usefulness of these models beyond their original purpose of describing and predicting the behavior of cellular networks becomes clearer, more systems biologists will be persuaded to make their models publicly available using the SBML standard required by the Biomodels database.

This work may usefully be extended to thinking about the effects of external pressures on protein networks. For example, chemotherapy drugs operate by interfering with one or more pathways in rapidly proliferating cells. In doing so they exert tremendous selective pressure on these highly mutagenic cells and as a result tumors frequently evolve resistance to chemotherapeutics. It may be possible to use techniques developed here to try to identify drug targets which operate on proteins whose reactions are more resistant to evolution, extending the efficacy of those drugs.

All data and computer code developed in the course of this research will be made publicly available.

References

[1] Judith A Blake, Carol J Bult, James A Kadin, Joel E Richardson, and Janan T Eppig. The Mouse Genome Database (MGD): premier model organism resource for mammalian genomics and genetics. *Database*, 39(November 2010):842–848, 2011.

[2] Kevin R Brown and Igor Jurisica. Databases and ontologies Online Predicted Human Interaction Database. *Bioinformatics*, 21(9):2076–2082, 2005.

[3] D Allan Drummond, Jesse D Bloom, Christoph Adami, Claus O Wilke, and Frances H Arnold. Why highly expressed proteins evolve slowly. *Proceedings of the National Academy of Sciences of the United States of America*, 102(40):14338–43, October 2005.

[4] D Allan Drummond, Alpan Raval, and Claus O Wilke. A Single Determinant Dominates the Rate of Yeast Protein Evolution. *Molecular Biology*, (2002), 2005.

[5] D Allan Drummond and Claus O Wilke. Theory Mistranslation-Induced Protein Misfolding as a Dominant Constraint on Coding-Sequence Evolution. *Cell*, pages 341–352, 2008.

[6] H.B. Fraser, A.E. Hirsh, L.M. Steinmetz, C. Scharfe, and M.W. Feldman. Evolutionary rate in the protein interaction network. *Science*, 296(5568):750, 2002.

[7] K. A. Geiler-samerotte, M. F. Dion, B. A. Budnik, S. M. Wang, and D. L. Hartl. Misfolded proteins impose a dosage-dependent fitness cost and trigger a cytosolic unfolded protein response in yeast. *PNAS*, 2010.

[8] Jodi Gureasko, Olga Kuchment, Debora Lika Makino, Holger Sondermann, Dafna Bar-Sagi, and John Kuriyan. Role of the histone domain in the autoinhibition and activation of the Ras activator Son of Sevenless. *Proceedings of the National Academy of Sciences of the United States of America*, 107(8):3430–5, February 2010.

[9] Ryan N. Gutenkunst, Jordan C. Atlas, Fergal P Casey, Robert S. Kuczenski, Joshua J Waterfall, Christopher R Myers, and James P Sethna. SloppyCell, 2007.

[10] Aric A Hagberg, Daniel A Schult, and Pieter J Swart. Exploring Network Structure , Dynamics , and Function using NetworkX. *Network*, (SciPy):11–15, 2008.

[11] Matthew W. Hahn and Andrew D. Kern. Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. *Molecular Biology and Evolution*, 22(4):803–806, 2005.

[12] Aaron E Hirsh and Hunter B Fraser. Protein dispensability and rate of evolution. *Nature*, 411(June):2–5, 2001.

[13] Motoo Kimura. Molecular evolutionary clock and the neutral theory. *Journal of Molecular Evolution*, pages 24–33, 1987.

[14] Chen Li, Marco Donizelli, Nicolas Rodriguez, Harish Dharuri, Lukas Endler, Vijayalakshmi Chelliah, Lu Li, Enuo He, Arnaud Henry, Melanie I Stefan, Jacky L Snoep, Michael Hucka, Nicolas Le Novère, and Camille Laibe. BioModels Database : An enhanced , curated and annotated resource for published quantitative kinetic models. *Systems Biology*, 2010.

[15] Ben-yang Liao, Nicole M Scott, and Jianzhi Zhang. Impacts of Gene Essentiality , Expression Pattern , and Gene Compactness on the Evolutionary Rate of Mammalian Proteins. *Gene*, 2005.

[16] Ben-Yang Liao, Nicole M Scott, and Jianzhi Zhang. Impacts of gene essentiality, expression pattern, and gene compactness on the evolutionary rate of mammalian proteins. *Molecular biology and evolution*, 23(11):2072–80, November 2006.

[17] Brian K. Mannakee and Ryan N. Gutenkunst. DynEvol, 2012.

[18] James O. McInerney. The causes of protein evolutionary rate variation. *Trends in Ecology and Evolution*, 21(5):230 – 232, 2006.

[19] Christopher R Myers and Ryan N Gutenkunst. Python Unleashed on Systems Biology. *Computing in Science and Engineering*, 9(34):1–4, 2007.

[20] Csaba Pál, Balázs Papp, and Martin J Lercher. An integrated view of protein evolution. *Nature reviews. Genetics*, 7(5):337–48, May 2006.

[21] Andrew I Su, Tim Wiltshire, Serge Batalov, Hilmar Lapp, Keith A Ching, David Block, Jie Zhang, Richard Soden, Mimi Hayakawa, Gabriel Kreiman, Michael P Cooke, John R Walker, and John B Hogenesch. A gene atlas of the mouse and human protein-encoding transcriptomes. *PNAS*, 2004.

[22] Alexander E Vinogradov. Systemic factors dominate mammal protein evolution. *October*, (January):1403–1408, 2010.

[23] Guang-Zhong Wang and Martin J. Lercher. The effects of network neighbours on protein evolution. *PLoS ONE*, 6(4):e18288, 04 2011.

[24] Ziheng Yang. PAML 4 : Phylogenetic Analysis by Maximum Likelihood. *Molecular Biology*, 24(Ml):1586–1591, 2007.

[25] Emile Zuckerkandl and L. Pauling. Evolutionary Divergence and Convergence in Proteins . *Evolving genes and proteins*, pages 97–166, 1965.

Legends, Tables, and Figures

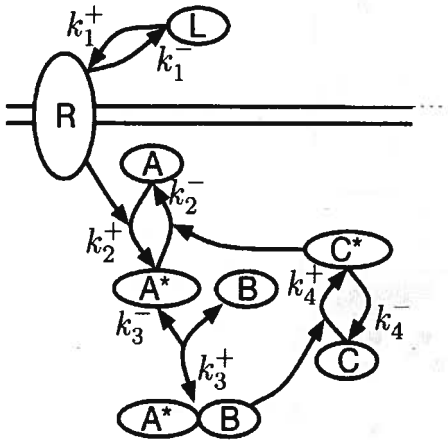


Figure 1: Signalling network

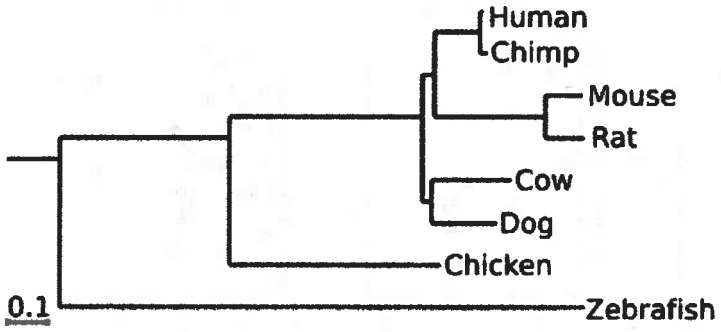


Figure 2: Vertebrate evolutionary tree

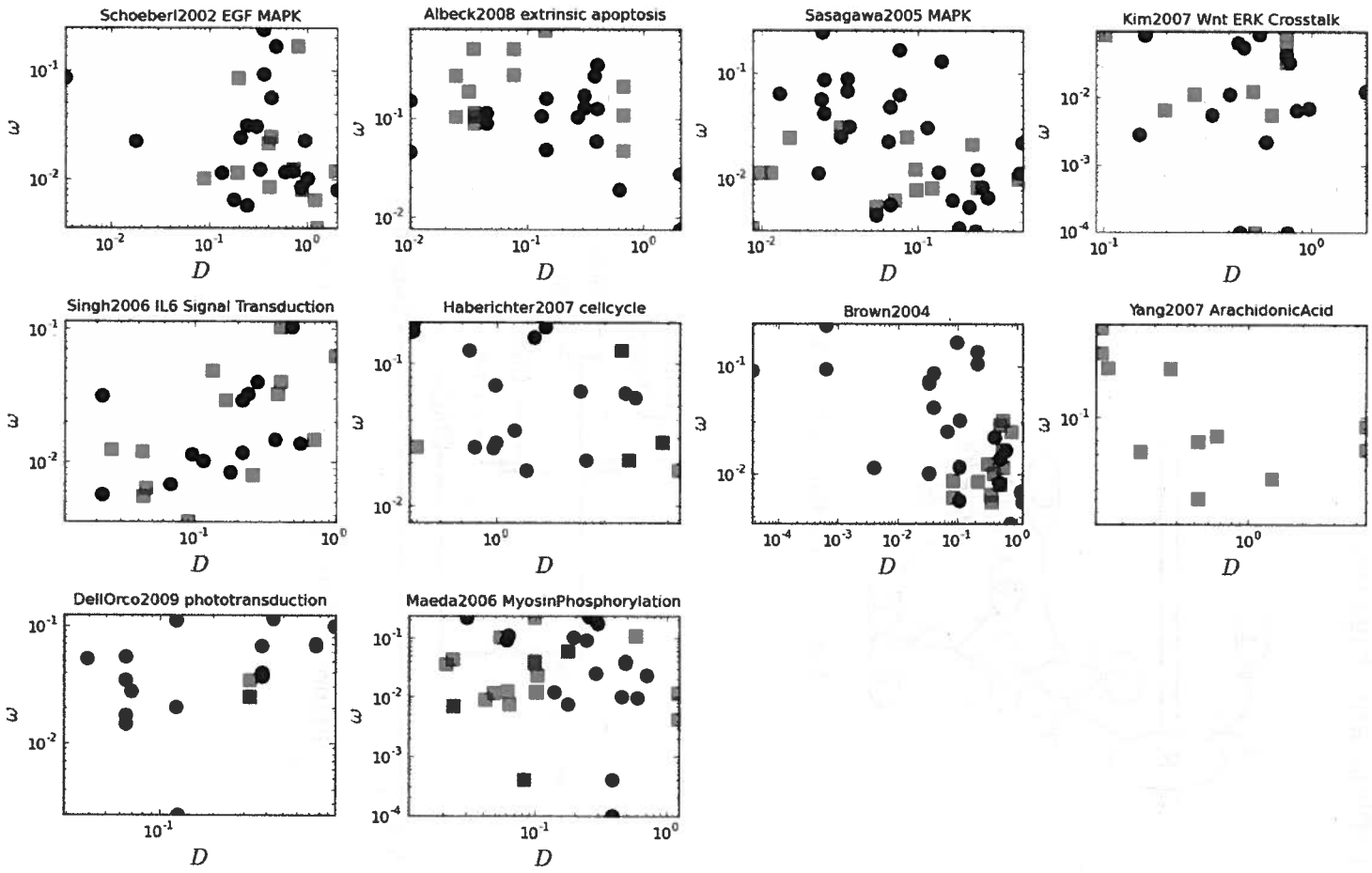


Figure 3: Log scale plot of Dynamical Influence D vs. Evolutionary rate ω for 10 cellular models. Each black dot represents a non-catalytic domain and each gray square represents a non-catalytic domain.

.

Table 1: Spearman rank correlations for evolutionary rate ω, dynamical influence D, and expression level X for all reaction types

Model	$\rho_{D,\omega}$	$\rho_{D\|X,\omega}$	$\rho_{X,D}$	$\rho_{X,\omega}$
Albeck2008 extrinsic apoptosis	-0.28	-0.31	-0.2	-0.11
Brown2004	-0.5	-0.48	0.18	-0.23
DellOrco2009 phototransduction	0.42	0.43	0.01	0.26
Haberichter2007 cellcycle	-0.24	-0.24	0.1	-0.01
Kim2007 Wnt ERK Crosstalk	-0.08	-0.25	-0.48	-0.28
Maeda2006 MyosinPhosphorylation	-0.12	-0.12	-0.2	-0.02
Sasagawa2005 MAPK	-0.35	-0.39	-0.17	-0.18
Schoeberl2002 EGF MAPK	-0.36	-0.26	0.39	-0.35
Singh2006 IL6 Signal Transduction	0.56	0.59	0.33	0.01
Yang2007 ArachidonicAcid	-0.59	-0.64	0.43	-0.03

Table 2: Spearman rank correlations for evolutionary rate ω, dynamical influence D, and expression level X for catalytic reactions

Model	$\rho_{D,\omega}$	$\rho_{D\|X,\omega}$	$\rho_{X,D}$	$\rho_{X,\omega}$
Albeck2008 extrinsic apoptosis	-0.01	0.07	0.17	-0.44
Brown2004	0.64	0.7	0.03	-0.38
Haberichter2007 cellcycle	-0.5	-0.53	0.1	0.2
Kim2007 Wnt ERK Crosstalk	0.24	-0.09	-0.75	-0.39
Maeda2006 MyosinPhosphorylation	0.09	0.11	-0.34	0.03
Sasagawa2005 MAPK	0.02	-0.06	-0.44	-0.17
Schoeberl2002 EGF MAPK	-0.33	-0.2	0.55	-0.3
Singh2006 IL6 Signal Transduction	0.65	0.72	0.33	-0.07
Yang2007 ArachidonicAcid	-0.59	-0.64	0.43	-0.03

Table 3: Spearman rank correlations for evolutionary rate ω, dynamical influence D, and expression level X for non-catalytic reactions

Model	$\rho_{D,\omega}$	$\rho_{D\|X,\omega}$	$\rho_{X,D}$	$\rho_{X,\omega}$
Albeck2008 extrinsic apoptosis	-0.16	-0.15	-0.47	0.07
Brown2004	-0.51	-0.48	0.18	-0.33
DellOrco2009 phototransduction	0.44	0.44	0.02	0.25
Haberichter2007 cellcycle	-0.28	-0.28	-0	-0.01
Kim2007 Wnt ERK Crosstalk	-0.07	-0.16	-0.28	-0.28
Maeda2006 MyosinPhosphorylation	-0.44	-0.45	0.18	-0.01
Sasagawa2005 MAPK	-0.54	-0.53	0.17	-0.18
Schoeberl2002 EGF MAPK	-0.15	-0.11	0.15	-0.35
Singh2006 IL6 Signal Transduction	0.61	0.6	0.27	0.14

Table 4: Principal component analysis

		Loadings as percentage of component				
Comp.	% Variance Expl.	Influence	Expression	Centrality	Degree	Specificity
1	47	0	0	90	10	0
2	31	2	95	0	1.5	1.5
3	13	97	3	0	0	0
4	8	0	2	8	90	0
5	1	0	1	0	0	99