Protein Sequence, Structure, and Dynamics Reveal Insights in the Divergence of Protein Functions

Okinawa Institute of Science and Technology
Graduate University

Thesis submitted for the degree
Doctor of Philosophy

Protein Sequence, Structure, and Dynamics Reveal Insights in the Divergence of Protein Functions

by
Stefano Pascarelli

Supervisor: **Paola Laurino**

April 2023

Declaration of Original and Sole Authorship

I, Stefano Pascarelli, declare that this thesis entitled "*Protein sequence, structure, and dynamics reveal insights in the divergence of protein functions*" and the data presented in it are original and my own work.

I confirm that:

- No part of this work has previously been submitted for a degree at this or any other university.
- References to the work of others have been clearly acknowledged. Quotations from the work of others have been clearly indicated, and attributed to them.
- In cases where others have contributed to part of this work, such contribution has been clearly acknowledged and distinguished from my own work.
- None of this work has been previously published elsewhere, with the exception of the following:

1. **Pascarelli S**, Merzhakupova D, Uechi G-I, Laurino P. Binding of single-mutant epidermal growth factor (EGF) ligands alters the stability of the EGF receptor dimer and promotes growth signaling. Journal of Biological Chemistry. 2021;297(1).

2. **Pascarelli S**, Laurino P. Inter-paralog amino acid inversion events in large phylogenies of duplicated proteins. PLoS computational biology. 2022;18(4):e1010016.

3. Toledo-Patiño S, **Pascarelli S**, Uechi GI, Laurino P. Insertions and deletions mediated functional divergence of Rossmann fold enzymes. Proc Natl Acad Sci U S A. 2022;119(48):e2207965119. Epub 2022/11/24. doi: 10.1073/pnas.2207965119. PubMed PMID: 36417431.

4. Dindo M, **Pascarelli S**, Chiasserini D, Grottelli S, Costantini C, Uechi GI, Giardina G, Laurino P, Cellini B. Structural dynamics shape the fitness window of alanine: glyoxylate aminotransferase. Protein Sci. 2022;31(5):e4303.

Date: April 18th, 2023

Signature:

Abstract

Proteins participate in every important aspect of known living systems. The amino acid sequence of a protein contains information about the physicochemical properties, the three-dimensional structure, and its function. However, connecting protein sequence to function is still an open challenge, particularly for protein families with many intramolecular interactions and therefore multiple functions. In this thesis, I show how I got insights into protein function using its evolutionary history, previous biological annotations, and Molecular Dynamics (MD) simulations. I developed two methods, one that relies on the ortholog conjecture and one that does not, and I used them to obtain mechanistic details on the Epidermal Growth Factor Receptor (EGFR), an important protein involved in development and cancer. By employing mutated ligands of EGFR, I found that these ligands are able to activate the receptor in alternative ways by affecting the stability of the dimerization interface of the receptor, therefore resulting in different downstream pathway activations. On another note, after looking at the evolutionary history of EGFR gene duplication, I discovered a pattern of protein sequence evolution that is related to a a special case of functional divergence between paralogs herein named “Meta-functionalization”. Then, I extended my analysis to more duplicated proteins, showing that Meta-functionalization could be a common mechanism for paralogs functional diversification. My software provided a way to identify the occurrence of this event and the residues responsible for it. In addition to EGFR, I investigated the dynamics of two proteins, scDH and hAGT – respectively, a Rossman fold dehydrogenases and the human Alanine-glyoxylate aminotransferases– by performing MD simulations. In this last part, my data extends on the experimental knowledge and provides mechanistic insights on the functional transition of these protein mutants and variants. Overall, my thesis shows a deep interconnection between functional divergence and protein sequence, structure, and dynamics, that can be exploited for the prediction of functional residues or the identification of evolutionary events. The conceptual foundations of this study could be used in other fields where gene duplication and functional residues play an important part, as for example in the search of mutants with alternative functions in protein engineering, and in the study of oncogene evolution after copy number variation in cancer biology.

Acknowledgements

First and foremost, my eternal gratitude goes to Professor Paola Laurino for her everlasting support, even in the darkest moments of my PhD journey. I want to thank my thesis committee members, Professors Timothy Ravasi and Evan Economo, for monitoring my progress and providing useful feedback. I want to mention Professor Federica di Palma, for hosting me in Norwich. A great effort was also made by my thesis proposal external examiner, Professor Nir Ben Tal, whose in-depth feedback of my thesis projects was deeply appreciated.
Next, I acknowledge all past and present members of my lab, Protein Engineering and Evolution Unit, for creating a friendly and stimulating work environment. I extend my sincere gratitude to all my paper co-authors: Barbara Cellini, Giorgio Giardina, Gen-Ichiro Uechi, Claudio Costantini, Silvia Grottelli, Davide Chiasserini, Dalmira Merzhakupova, and in particular Dr. Saacnicteh Toledo-Patino, and Mirco Dindo for leading the efforts of our collaborations. The work in this thesis would have never been possible without the support of the Scientific Computing and Data Analysis (SCDA) section and Jan Moren, that handle the OIST supercomputers. I would like to extend my gratitude to OIST admins and staff, especially Graduate School, who always try their best to make sure that everything goes smoothly.
My family was extremely supportive of my long journey to the other side of the world, for this I will be forever grateful. Thank you, Rosa, Antonio, and Davide, for bearing with my absence. My biggest motivator, Christine, who I can't thank enough for her constant presence and helpful scientific discussions. I also want to thank all my friends in Italy and those in Japan, for keeping up with me. Special mentions in random order go to Tabbal, Miles, Akira, Julian, Giuseppe, Nicola, Alessandro, Sam, Soumen, Silvia, Leonardo, Mirco, Riccardo, Paolone, Ianto, Lorena, etc. I thank every person who contributed to the OIST football club and my frisbee club, so that I could keep a *mens sana in corpore sano*. Lastly, thanks for all the support to the student council, in which I volunteered my time to fight for the students' rights. Thanks to all these experiences, I was able to learn much more than science during my PhD.

Abbreviations

AGT-Ma/Mi, Alanine:Glyoxylate aminotransferase Major/Minor allele; **AREG**, Amphiregulin; **BTC**, Betacellulin; **CASP**, Critical Assessment of protein Structure Prediction; **CD**, circular dichroism; **DCA**, Direct Coupling Analysis; **DDC**, Duplication-Degeneration-Complementation; **DMEM**, Dulbecco's Modified Eagle Medium; **DNA**, DeoxyriboNucleic Acid; **EC**, Enzyme Commission; **ECD**, Extra Cellular Domain; **ECOD**, Evolutionary Classification Of protein Domains; **EGF**, Epidermal Growth Factor; **EGFR**, Epidermal Growth Factor Receptor; **ENA**, European Nucleotide Archive; **EPGN**, Epigen; **EREG**, Epiregulin; **ET**, Evolutionary Trace; **FAD**, Flavin Adenine Dinucleotide; **FES**, Free Energy Surface; **FRET**, Fluorescence Resonance Energy Transfer; **GO**, Gene Ontology; **HBEGF**, Heparin-Binding Epidermal Growth Factor; **HMM**, Hidden Markov Model; **IAD**, Innovation Amplification Divergence; **ITC**, Isothermal Titration Calorimetry; **IDP**, Intrinsically Disordered Proteins; **LC/MS**, Liquid Chromatography/Mass Spectrometry; **MC**, Marginally Conserved; **MD**, Molecular Dynamics; **MSA**, Multiple Sequence Alignment; **MST**, MicroScale thermophoresis; **MSTA**, Multiple STructural Alignment; **NAD**, Nicotinamide Adenine Dinucleotide; **NMR**, Nuclear Magnetic Resonance; **PDB**, Protein Data Bank; **PH1**, Primary Hyperoxaluria type 1; **PLP**, PyridoxaL Phosphate; **PPI**, Protein-Protein Interaction; **QM/MM**, Quantum Mechanics/Molecular Mechanics; **RMSD**, Root Mean Square Deviation; **RMSF**, Root Mean Square Fluctuation; **RNA**, RiboNucleic Acid; **SAM**, S-Adenosyl Methionine; **scDH**, short chain dehydrogenase/reductase; **SCOP**, Structural Classification Of Proteins; **SDS**, Specificity Determining Sites; **SSN**, Sequence Similarity Network; **TSGD**, Teleost Specific whole Genome Duplication; **WGD**, Whole Genome Duplication; **WT**, Wild Type;

List of Publications

Publications and author contributions are listed in the order of appearance in the body of this thesis:

1. **Pascarelli S**, Merzhakupova D, Uechi G-I, Laurino P. Binding of single-mutant epidermal growth factor (EGF) ligands alters the stability of the EGF receptor dimer and promotes growth signaling. Journal of Biological Chemistry. 2021;297(1).

Contributions
PL, SP, DM conceived and designed the project; SP developed and performed the computational models and analyses; DM and GU performed the experiments; all authors analyzed the data and wrote the manuscript.

2. **Pascarelli S**, Laurino P. Inter-paralog amino acid inversion events in large phylogenies of duplicated proteins. PLoS computational biology. 2022;18(4):e1010016.

Contributions
SP performed the formal analysis and validation, developed the software, curated the data and the visualization. PL was in charge of the project administration, supervision, and funding acquisition. All authors conceptualized the work and wrote the manuscript.

3. Toledo-Patiño S, **Pascarelli S**, Uechi GI, Laurino P. Insertions and deletions mediated functional divergence of Rossmann fold enzymes. Proc Natl Acad Sci U S A. 2022;119(48):e2207965119. Epub 2022/11/24. doi: 10.1073/pnas.2207965119. PubMed PMID: 36417431.

Contributions
S.T.P. and P.L conceived the project. S.T.P. designed the bioinformatical pipeline for sequence and structural analyses, expressed and purified proteins for structure solving. G.U. and S.T.P. conducted protein expression and their binding analysis with ITC. S.P. performed the molecular dynamics simulations. S.T.P., S.P., and P.L. analyzed the data. S.T.P. and P.L. wrote the manuscript with inputs from S.P. This project was supervised by PL.

4. Dindo M, **Pascarelli S**, Chiasserini D, Grottelli S, Costantini C, Uechi GI, Giardina G, Laurino P, Cellini B. Structural dynamics shape the fitness window of alanine: glyoxylate aminotransferase. Protein Sci. 2022;31(5):e4303.

Contributions
B.C., G.G., P.L. and M.D. conceived the project. M.D. and P.L. planned the library. M.D. and G.U. generated the library and tested the activity. M.D. expressed, purified and characterised single mutants. S.P. performed molecular dynamics simulations, rate of evolution and phylogenetic analysis. G.G. collected crystallography data, processed, solved and analysed crystallography data. S.G. and D.C. performed experiments in Hek293 cells and analysed MS data. C.C. analyzed Hek293 and MS data and revised the manuscript. B.C., P.L. and G.G. supervised the project and wrote the manuscript with the input from all the authors.

Dedication

To the memory of nonna Regina Maria De Vita (27th November 1931- 4th January 2018), without whom I would have never made it this far

Table of Contents

Introduction

Living systems are patchworks of repurposed parts miraculously working together. The logic behind their formation is simple, as evolution works by just filtering random variability. However, the outcomes of this process are so astonishing that they can trick even the eyes of the experts into thinking of a greater plan behind it. For example, we observed wings and flippers appear in bats and seals, co-opted from primordial limbs, to serve their specific purpose. At the microscopic scale, the changes that we see might be less evident, but their results can be as dramatic. Adding one single human gene in a chimpanzee brain organoid increased the number of basal progenitor cells to be human-like (Fischer et al., 2022), or a human protein fusion might induce terminal diseases, as in the Philadelphia chromosome leukemia (Nowell & Hungerford, 1960). We are a product of evolution. Therefore, understanding its mechanism would answer the great questions of how we came to exist and why.

Evolution takes place at multiple levels (genes, cells, organisms, populations) and with different subjects (living things, languages, memes). However, a good starting point to study it would be the most basal subject: genes and their encoded proteins. The raw material of evolution is found at the central dogma of molecular biology, where DNA, RNA, and proteins are tightly intertwined. This system is already a complex one; how it arose is still one of the biggest mysteries, as we cannot yet decompose it in stable, self-replicating smaller parts (Alberts et al., 2008). There are many unknowns regarding the status quo in Biology and how it became the norm. Thus, a study of evolution at the molecular level is now overdue.

Proteins

Proteins are the molecular effectors of the cell, taking part in many important processes for sustaining life at the most basal level. A protein is generally made by combinations of one or more domains, independently foldable units tightly connected to a specific function. The intrinsic modular structure of proteins led to a “biological big bang” (Dokholyan, Shakhnovich, & Shakhnovich, 2002) that geared up the range of possible protein structures, by mechanisms still unknown (Moore, Björklund, Ekman, Bornberg-Bauer, & Elofsson, 2008). However, testing a hypothesis in such a remote past using the current methods is challenging. The protein world is extremely non-homogeneous, limited by functional and mechanical constrains (Koonin, Wolf, & Karev, 2002).. To capture the diversity of protein functions, several classifications have been proposed. In a straightforward manner, the enzyme commission (EC) number categorize every protein according to the chemical reaction that they catalyze (Barrett, 1997). While this classification provides useful data on the metabolic network structure of an organism, EC numbers are restricted to the biochemical function, thus excluding most proteins involved in other functions. Instead, the Gene Ontology (GO) consortium proposed a relational representation of protein function divided in three main aspects: the molecular function, the cellular component, and the biological process (Ashburner et al., 2000). This general classification takes into account the contextual nature of protein function, where multi-functionality, or “Moonlighting”, is a common feature that confound protein annotation (Constance J. Jeffery, 1999). Proteins with multiple functions are informally referred as moonlighting, from moonlighting workers, having an additional job (Copley, 2012; C. J. Jeffery, 2003). The relevance of moonlighting proteins has been deeply analyzed by Piatigorsky in his book ‘Gene sharing and Evolution’, mainly focusing on protein crystallin (Piatigorsky, 2007).

The three aspects of protein function depend on a plethora of factors, including protein-protein interactions, post-translational modifications (e.g., phosphorylation and methylation), compartmentalization, complex formation, and so on. However, the most fundamental factor that determines protein function is protein structure and its dynamics (Hensen et al., 2012). For example, enzymes usually fold into pockets that can bind the substrates and move to release the products. The knowledge of protein structure can be divided into four levels, from primary to quaternary. At the very base, primary structure, or protein sequence, is encoded by the DNA, thus providing a direct connection between gene mutations and function. This simple representation by a sequence of characters hides a multi-layer depth of knowledge, derived from the sequence itself and from the comparison to other existing proteins.

Primary structure

The most abundant data of the protein world is sequences. In recent decades, protein sequences have been obtained as a byproduct of genomic data, with a higher rate than any other characterization of individual proteins. The reason behind it is the effort by the international scientific community towards obtaining as many genomic (Hotaling, Kelley, & Frandsen, 2021) and metagenomic (L. R. Thompson et al., 2017) DNA sequences as possible. UniRef90, an online resource collecting non-redundant protein sequences (Suzek, Huang, McGarvey, Mazumder, & Wu, 2007), has about 10^8 entries, at least three orders of magnitude more than the databases for protein structure, the PDB (Berman et al., 2000), and protein function, the Gene Ontology database (Ashburner et al., 2000; Thomas et al., 2022).

There is a long history of protein-sequence-based bioinformatics. From the 70s onward, scientist tried to answer important biological questions about proteins despite the limited computational resources and data available. At that time, protein folding was one of the main topics of discussion. The mindset driving the scientific efforts in that period is hinted by the Nobel winning work behind Anfinsen's dogma (Anfinsen, 1973). Anfinsen stated that a protein sequence is enough to determine its three-dimensional structure. From then on, a concerted effort was spent on developing more and more refined computational methods to effectively predict protein structure. Starting from 1994, the best methods have been tested biannually in the CASP challenge (Pereira et al., 2021). CASP, or Critical Assessment of Protein Structure, is a double-blind study where several research groups submit their developed algorithms to predict the unknown structure of proteins. During the years, the community steadily progressed, until one method that particularly excelled was found in CASP14. Fifty years after the Anfinsen's dogma, AlphaFold2 (Jumper et al., 2021) provided a general solution to the protein folding problem that will inevitably extend the potential of the simple protein sequence. Consistently predicting the tertiary structure of a protein from its sequence is a paradigm shift in Bioinformatics that produced an immediate impact, while paving the way for many more applications yet to come.

In alternative to structure prediction, protein sequences have been used to define and identify protein domains. Thanks to the accumulation of sequences in biological repositories, scientists were able to identify patterns of similarity between parts of them, in what has been named protein domain. Domains are units of conserved sequences (Schaeffer & Daggett, 2011) that are usually related to independently folding units sharing a particular function (Finn et al., 2008; Marchler-Bauer et al., 2007). The presence and architecture of domains in a protein sequence can be obtained with several methods (Y. Wang, Zhang, Zhong, & Xue, 2021), usually involving Hidden Markov Models (Eddy, Mitchison, & Durbin, 1995; Remmert, Biegert, Hauser, & Söding, 2012), statistical models representing the alternative sequences that a domain can be found with. The identification of domains on a protein

sequence is referred to as “Annotating” the sequence, and it often provides indications about the general function of the protein (Rojano et al., 2022). The significant development of domain annotation was built on top of the efficient global and local alignment algorithms developed in the early stages of Bioinformatics (Altschul, Gish, Miller, Myers, & Lipman, 1990; Higgins, Thompson, & Gibson, 1996; Needleman & Wunsch, 1970). Since that time, it was clear that the repeated patterns observed in protein sequences would be critical to understand the protein world, by the use of a measure that could let scientists glance into the world of primitive amino acid sequences: sequence conservation (Eck & Dayhoff, 1966).

Sequence conservation

Protein sequence contains much more information than the list of its amino acids. The sequence can be used to predict physicochemical properties of the amino acid chain, like the acid dissociation constant (pK_a), molecular weight (Mw), or solubility (Oeller, Kang, Sormanni, & Vendruscolo, 2022). The amino acid composition was even used as a measure of similarity between proteins (Yu, Zhang, Gutman, Shi, & Dehmer, 2017) to obtain an alignment-free measure with comparable results to competitors such as Clustal W (J. D. Thompson, Higgins, & Gibson, 1994). Though, a further layer of information was extracted by moving from the analysis of a single protein sequence to the comparison of similar proteins observed in public databases. Initially, the comparison of similar protein sequences led to the development of PAM (Dayhoff M, 1972) and BLOSUM (Henikoff & Henikoff, 1992) matrices, a statistical model of the observed amino acid substitutions. These matrices quantify the observations regarding variable positions in a protein alignment at discrete similarity thresholds, using empirical data. The expected value of each amino acid substitution worked well as a background distribution and was later used to establish evolutionary distance and phylogenetic relationships (Sonnhammer & Hollich, 2005).

Not only what is observed, but also what is not observed is an exploitable piece of information. This is the concept behind the measure of sequence conservation. Considering a protein Multiple Sequence Alignment (MSA), where the sequence of similar proteins is aligned to find corresponding positions, some columns might be found without a change of amino acid. These sites usually represent a position in the protein that has a critical function; for example, one in charge of the catalysis in the substrate binding pocket of an enzyme. In these cases, the position is said to be conserved and it likely means that a variation of amino acid is not favorable. In fact, sequence conservation was observed to be inversely correlated with tolerance to mutation (Guo, Choe, & Loeb, 2004). In the years, several methods have been developed to quantify sequence conservation. One of the first and commonly used method calculates the Shannon entropy of columns in the MSA (Durbin et al., 1998), but other more elaborate approaches make use of Von Neumann entropy (Caffrey, Somaroo, Hughes, Mintseris, & Huang, 2004), Bayesian phylogenetic tree evolutionary rates (Ashkenazy, Erez, Martz, Pupko, & Ben-Tal, 2010; Mayrose, Graur, Ben-Tal, & Pupko, 2004), or Jensen-Shannon divergence (Capra & Singh, 2007). The work on conservation was used as a foundation for more specific analyses of protein function, including but not limited to co-evolution measures (de Juan, Pazos, & Valencia, 2013), contact predictions (Skwark, Abdel-Rehim, & Elofsson, 2013), functional residues (discussed later) and protein structure predictions (Jumper et al., 2021).

Homology and duplication

The protein sequences we observe today are the product of million years of refinement through evolution. Just like in human society, two proteins might be related to each other by

sharing a common ancestor. The measure of sequence similarity, or conservation, between them might reflect their degree of relatedness. Proteins that share a common evolutionary origin are called homologs. When they are found in different organisms, if they correspond to the same protein in the progenitor, they are called orthologs (Fitch, 1970). The identification of orthologs is essential in many phylogenetic analyses and comparative genomics (Gabaldón & Koonin, 2013), but it is still an open problem in Bioinformatics. The underlying mechanism governing protein evolution makes the ortholog relationship difficult to detect. Positive selection increases the rate of mutation observed on the DNA, reducing the degree of conservation between homologs. When the sequence identity, a rough measure of conservation, is below 30%, the homology becomes more challenging to predict with current methodologies. The area below this threshold is commonly known as the "Twilight zone" of sequence similarity (Rost, 1999). Additionally, rearrangements at the chromosome level could generate hybrid proteins with fused or recombined domains (Björklund, Ekman, Light, Frey-Skött, & Elofsson, 2005). When a genetic mechanism generates two copies of a gene, two identical proteins start to diverge semi-independently. If the two genes were generated in a Whole Genome Duplication (WGD) event, the two copies are called ohnologs, else the general term is paralogues. A notable effort is ongoing to correctly identify orthologs and paralogs among false positive protein sequences (Altenhoff et al., 2020). Searching for the most similar annotated ortholog is also a common practice when studying an unknown protein (Loewenstein et al., 2009). The underlying assumption, named as "The Ortholog Conjecture", is that orthologs are more likely to share the same function rather than paralogs. This hypothesis generated a heated debate (Chen & Zhang, 2012; Nehrt, Clark, Radivojac, & Hahn, 2011; Stamboulian, Guerrero, Hahn, & Radivojac, 2020), showing that there is a missing gap in our knowledge of protein homology. By addressing it, we might uncover deep insights into the evolution of proteins and possibly of the mechanism of evolution in general.

Duplication and function

"Natural selection merely modified, while redundancy created" (Ohno, 1970). With this sentence, Susumu Ohno underlined the importance of gene duplication as a generator of novelty. Ohno is the first to compile a broad perspective of gene duplication in his fore sighting work: "Evolution by gene duplication", in 1970. First, he argues about the conservativeness of natural selection. As later confirmed by studies of protein fitness landscape, the freedom of a protein mutating from one functional optimum to another one is limited by the loss in fitness of the transition steps. When this transition happens, it was observed to involve usually a 'generalist' intermediate protein, that is able to carry out multiple functions at the same time (Levin et al., 2009). Ohno makes a point by arguing that duplication overcomes this constrain by providing a fresh template that could be modified multiple times, as long as the spare copy is still carrying out the original function. Two classic examples of this process are the trypsin/chymotrypsin (Baptista, Jonson, Hough, & Petersen, 1998; McLachlan, 1979) and the hemoglobin (Storz, 2016). Both proteins had a similar fate. The duplication and subsequent divergence from an ancestral protein led to new activities. Ohno's views on duplication are fascinating, still they leave some questions unanswered. Foremost, how a new function emerges from a duplicated gene. Knowing this would have a cascade effect on answering how duplicated genes get fixed inside a population, and how the initial redundancy is maintained in spite of the risk of pseudogenization and other shortcomings of gene duplication.

More recent theories have stressed the importance of sub-functionalization in the protein duplication context (Rastogi & Liberles, 2005). Sub-functionalization happens when

the two paralogs retain only a subset of the functions of the original protein. The two paralogs are therefore able to specialize one function, without the risk of compromising other functions now carried out solely by the second copy. Interestingly, this process does not require a selective pressure to take place. A neutral mutation setting gives a "more parsimonious" explanation of duplication retention in genomes. This model is referred in literature as the Duplication-Degeneration-Complementation (DDC) model (Force et al., 1999). The relationship between sub- and neo-functionalization is still hotly debated. Work on duplicated HOX genes expression in zebrafish seems to support the DDC model, though modelling work by Rastogi and Liberles shows how sub- functionalization is restricted to the first phase after duplication, while neo-functionalization has an everlasting presence with time (Rastogi & Liberles, 2005). Another work based on computational modeling showed that mutational robustness might also play a role in the balance of sub-functionalization outcomes (Sikosek, Chan, & Bornberg-Bauer, 2012). The concept of sub-functionalization can only be applied to multifunctional proteins. Though, the preponderance of multifunctional proteins can be inferred by the complexity of protein networks. An interesting picture of how the complex topology of protein networks may increase its reliability is found in Kitano 2009(Hase, Tanaka, Suzuki, Nakagawa, & Kitano, 2009). In the context of biological networks, protein duplication has been suggested as an important mechanism, both for increasing robustness and as a mechanism to generate a scale-free like distributions in biological networks (Hughes & Friedman, 2005). Protein connectivity and duplications have a further interesting implication. Highly connected "HUB" proteins have a slower rate of evolution compared to their interacting partners, at least in human (Alvarez-Ponce, Feyertag, & Chakraborty, 2017). Duplication of hub proteins often result in the uneven loss of protein interactions (Roth et al., 2007), though rearranging the connectivity of the two copies. This event might act as a release of the evolutionary constrains of the protein, from a definite time point initiator, the time of duplication. This gives an optimal source to study how evolution affects the protein at the sequence and functional level.

Function connected to sequence

The primary structure of a protein reflects its chemical composition. In theory, this information encodes much more than the 3D-structure. However, the underlying process that governs protein folding is still not completely understood (Li, Fooksa, Heinze, & Meiler, 2018). While artificial intelligence paved the way to solve this problem (Jumper et al., 2021; Senior et al., 2020), the tertiary structure is still not sufficient to unequivocally predict the function of all proteins. For this reason, several research groups have focused their efforts in creating algorithms for this purpose. Most approaches are based on amino acid sequence, combining available structural and evolutionary information (Nemoto, Saito, & Oikawa, 2013). A remarkable work was achieved in Evolutionary Trace (ET) (Lichtarge, Bourne, & Cohen, 1996), an algorithm that ranks amino acid residues in a protein sequence by their relative evolutionary importance. This pioneering work opened the road for more algorithms with a similar purpose. In ConSurf webserver (Ashkenazy et al., 2016), evolutionary rates of single amino acid are mapped from an input query protein, by several steps that involve homologs detection in protein databases, generation of a phylogenetic tree, and using advanced probabilistic evolutionary models (Pupko, Bell, Mayrose, Glaser, & Ben-Tal, 2002). A meaningful contribution to improve the accuracy of functional cluster prediction was obtained by the integration of Direct-Coupling Analysis (DCA) (Morcos et al., 2011), or other measures of co-evolution. Additionally, protein function could be tracked using Sequence Similarity Networks (SSNs) (Atkinson, Morris, Ferrin, & Babbitt, 2009). SSNs can compare hundred thousand of sequences in a computationally affordable way, by using

an all-vs-all distance matrix, usually representing sequence similarity. This approach was successfully employed to track SARS-CoV-2 mutations across US (Patil, Catanese, Brayton, Lofgren, & Gebremedhin, 2022). For a more complete collection of the existing methodologies for protein functional inference, refer to the review (Lee, Redfern, & Orengo, 2007).

A relevant number of published articles is devoted to the detection of Specificity Determining Sites (SDSs). SDS are positions that are involved in conferring a different function among subsets of a protein family homologs. Detection of SDSs is relevant for understanding the evolution of function after gene duplications and might be the base for protein engineering experiments and in silico evolution. SDS are the end product of the functional divergence of proteins after gene duplication, though their detection is hindered by the presence of neutrally evolving sites (Kimura, 1991). To date, three types of functional divergence have been described, depending on the evolutionary rate after duplication (Chakraborty & Chakrabarti, 2014). Type 1 sites resemble Ohno's view of duplication, where one of the two subfamilies show high degree of conservation, while the second subfamily presents a high evolutionary rate. In the case of type 2 divergence, high conservation is present in both subfamilies, preceded by an early differentiation that lowered the identity among the two. A third type of divergence was also identified: Marginally Conserved (MC) sites. For MC sites, no apparent sign of conservation is observed within any of the subfamilies (Chakrabarti, Bryant, & Panchenko, 2007).

In the last two decades, several computational algorithms have been developed for the detection of SDSs. These algorithms can be divided in five categories based on their mode of detection: entropy-based, evolutionary rate-based, automated subgrouping-based, 3D structure-based, machine learning- and feature-based. The methods appear to have a certain degree of complementarity. In fact, ensemble approaches that combined predictions made by different algorithms were are able to outperform each single method (Chakrabarti & Panchenko, 2009).

Function connected to structure and dynamics

Biological function is believed to be determined by the molecular structure of proteins (Leman et al., 2020). Following this principle, several methods have been developed to employ protein structure in a wide range of applications. Arguably, the most comprehensive collection is Rosetta, an impressive toolset of Bioinformatics software that can address structure prediction (Song et al., 2013), de-novo protein design (Jacobs et al., 2016), docking (Meiler & Baker, 2006), and much more. Typically, structural analysis is performed when there is a need to modify desired properties of a protein through mutations. Though, the complexity of the search grows exponentially with the length of the sequence. The number of possible proteins that differs for only one mutation from a starting sequence is 19^L, where L is the protein length. This number rapidly increases with more mutations. Therefore, one of the greatest challenges is to reduce the complexity of finding mutations that have a desired effect in a gigantic search space. When a structure is available, rational design is still one of the most frequent approaches. The manual curation of an expert can identify the residues responsible of specific biochemical properties, like thermostability, substrate specificity, and kinetic behavior (Pongsupasa, Anuwan, Maenpuen, & Wongnate, 2022). Depending on the application, rational design can have different approaches, ranging from purely combinatorial to highly rational, usually aided by design tools (Korendovych, 2018). Rational design can also be improved by a directed evolution approach (Yip et al., 2011). The directed evolution strategy consists of several rounds of simulated evolution, enacted by inducing mutations and successively selecting a desired property of the mutants (Bloom

& Arnold, 2009). Protein mutations are an efficient strategy to study the wild-type protein functions, as well as the effect of deleterious variants. These studies rely on the observation of a phenotype that will explain how the wild-type protein is disrupted when mutated in specific positions.

As previously described, the three-dimensional structure of proteins is commonly employed to track or understand protein function. However, this representation overlooks the fourth dimension: time. In fact, proteins are in a vibrating environment where movement and interactions are fundamental to carry out their functions. Protein dynamics becomes particularly relevant when the protein structure has multiple conformations or no conformation at all, as in Intrinsically Disordered Proteins (IDPs) (Dunker et al., 2001). IDP or IDP regions exhibit a biological activity even though having no stable structure (Oldfield & Dunker, 2014), and have been considered a big part of the dark proteome (Perdigão et al., 2015). It is clear that much more can be understood of protein functions by looking at time-dependent properties of proteins. Insights into protein dynamics can be obtained experimentally using for example Nuclear Magnetic Resonance (NMR) spectroscopy (Wüthrich, 2001) or Fluorescence Resonance Energy Transfer (FRET) spectroscopy (Mazal & Haran, 2019). An initial evaluation of the dynamics of the different parts of a protein can be already obtained from x-ray B-factors (Rueda et al., 2007), though with several limitations of applicability (Sun, Liu, Qu, Feng, & Reetz, 2019). Beside experimental methods, computational methods provide more freedom to investigate and alter a system of interest. Among all, molecular dynamics is a method that allows a detailed analysis of the system (M. Karplus & McCammon, 2002). Molecular Dynamics (MD) software like CHARMM (Brooks et al., 2009), AMBER (J. Wang, Wolf, Caldwell, Kollman, & Case, 2004) and GROMACS (Van Der Spoel et al., 2005) simulate Newton's equation of motion on full-atom models of proteins or other molecules. MD simulations have been successfully applied to uncover mechanistic details of how protein function, in particular concerning protein folding and enzyme catalysis (M. Karplus & Kuriyan, 2005). The software and hardware required to perform MD simulations has become more accessible and powerful during the years, leading to a growing attention by the scientific community (Hollingsworth & Dror, 2018) and pharmaceutical industry (De Vivo, Masetti, Bottegoni, & Cavalli, 2016). Regardless, full-atom classical MD has several disadvantages. Firstly, it requires a structure to start with, which in case of IDPs, multidomain, or transmembrane proteins could be challenging to obtain. The simulation requires very short time steps ($\sim 10^{-15}$s), resulting in a very high computational load when studying slower biological processes. Furthermore, in a classical MD, no covalent bonds form or break, denying the possibility to observe an enzymatic reaction taking place. To overcome these limitations, several modifications have been proposed. A type of mixed MD simulation combining quantum mechanics and molecular mechanics (QM/MM) has been successful in describing chemical reactions in a computational cost-efficient way (Senn & Thiel, 2009). Meanwhile, an increased sampling of conformations or time can be obtained with steered MD (Bernardi, Melo, & Schulten, 2015; Harpole & Delemotte, 2018), or coarse-grained simulations (Marrink & Tieleman, 2013). Overall, MD simulations generate a representation of protein dynamics that often reflects experimental characterization (Rueda et al., 2007). This view of the simulated molecular world allows us to investigate the closest layer to protein function, and it gives us an additional way to classify proteins using their range of motions, in the so called dynasome (Hensen et al., 2012).

The combination of protein sequence, structure, and dynamics data provides an opportunity to study protein function beyond the limits conveyed by the physical entity. In light of the recent technological developments, we can now explore protein function broadly through a phylogenetic tree, thus refining our understanding of the mechanism of evolution,

and its spontaneous generation of complexity. In the previous parts, I showed how the new methods that brought a technological advancement are built upon previous discoveries. Likewise, the scope of this work is twofold; to develop new methods based on the current methodologies, and to find new applications to the existing algorithms. Overall, the method development and data analysis that I performed in this work showcase an extension of what is achievable by protein bioinformatics. Though, the specifics of each chapter will be discussed accordingly in the following sections.

1. Chapter 1

Detecting Functional Divergence Among the Paralogous Ligands of the Epidermal Growth Factor Receptor

(Published in *Journal of Biological Chemistry* as *"Single EGF mutants unravel the mechanism for stabilization of Epidermal Growth Factor Receptor (EGFR) system"*)

1.1. Introduction

The Epidermal Growth Factor Receptor is a membrane-anchored receptor tyrosine kinase, member of the ErbB protein family. The fact that proteins in this family are involved in multiple cancer types has increased the interest in elucidating its mechanism of action and related disfunctions. EGFR is ubiquitously expressed, taking part in several physiological functions such as development, cell adhesion and migration, tissue regeneration, and others. In mice, the knock-out mutant suffers premature death, with abnormalities in multiple tissues. When transformed in oncogene, the dysregulated activation of EGFR can induce survival, proliferation, migration, growth and inhibition of apoptosis through the activation of a multitude of downstream pathways (Wee & Wang, 2017). The effects of EGFR activation are cell-type specific (Bjorkelund, Gedda, & Andersson, 2011; Johnson, Baxter, Vlodavsky, & Gospodarowicz, 1980) for yet unknown reasons. In human, the seven paralogous ligands of EGFR induce different pathways selectively with a mechanism that is independent of the ligand binding affinity or potency (Wilson, Gilmore, Foley, Lemmon, & Riese, 2009), in a process labeled as 'biased signaling' (Lane, May, Parton, Sexton, & Christopoulos, 2017). While much work has been done to study the effects of the binding of wild type ligands to the receptor, the effects of mutated ligands on the 'biased signaling' of EGFR is underexplored.

In this work, I developed a methodology to identify the functional residues involved in the paralog-specific function of EGFR ligands. The method relies on conservation and co-evolution measures based on the known evolutionary relationships of the paralogs, and their interaction to the receptor protein. Then, I analyzed the effects at the receptor and cell level of four modified EGF ligands that have been mutated at the high scoring positions. Although having comparable binding affinities, the EGF mutants induced a different level of phosphorylation of the receptor and growth rate in Bj5-tα fibroblasts. Finally, molecular dynamics showed that the binding of the mutant ligands had an effect on the dimerization interface of EGFR.

1.2. Published Article

Pascarelli S, Merzhakupova D, Uechi G-I, Laurino P. Binding of single-mutant epidermal growth factor (EGF) ligands alters the stability of the EGF receptor dimer and promotes growth signaling. *Journal of Biological Chemistry*. 2021;297(1).

1.3. Conclusions

In this chapter, I showed how the evolutionary context of a family of paralogs can be used to identify residues involved in paralog specific functions. The paralogous ligands of EGFR all share one function, the ability to bind the receptor. However, they induce different pathways by ‘biased signaling’. With the DIRPred method, I identified four positions that, when mutated in EGF, altered the ligand effect on the receptor. This study shows an interesting way investigate the mechanism of function of the EGF receptor, by mutating on the ligands, that has the potential to be applied in other protein-protein interaction systems.

2. Chapter 2

Protein Sequence Patterns of Evolution Signal Functional Transitions in Large Phylogenies

(Published in *PLoS computational biology* as *"Inter-paralog amino acid inversion events in large phylogenies of duplicated proteins"*)

2.1. Introduction

In this chapter, I shifted my focus on the EGF receptor to study the mechanisms of protein evolution after gene duplication. During vertebrate evolution, EGFR was included in the set of proteins that have been preserved after multiple rounds of genome duplication. The expansion of protein kinases in the genomic repertoire of higher vertebrate is considered one of the funding reasons of their complexity (Brunet, Volff, & Schartl, 2016). Phylogenetic studies suggest that one copy of EGFR was already present in the last common ancestor of the metazoans. Though, the expansion of EGFR family of RTK and ligands happened in the Chordata clade, as well as independently in few other Bilateria, like Platyhelminthes and Annelida (Barberán, Martín-Durán, & Cebrià, 2016). Interestingly, among Chordata, the conservation of EGFR shows alternating levels of amino acid similarity across its domains. High degree of conservation is found in the protein kinase domain (~90% similarity), while much lower conservation is observed when comparing the extra cellular domains (~60%). According to Laisney et al., this reflects a lineage specific co-evolution of the ECD with its ligands (Laisney, Braasch, Walter, Meierjohann, & Schartl, 2010). In particular, I chose to focus on the fish lineage for several contributing factors: 1) Ray-finned fish is the most successful vertebrate radiation event and therefore a powerful model group to focus on the evolution of the EGFR system (Volff, 2005); 2) The breadth of the ray-finned fish phylogeny offers a unique opportunity to study the EGFR system in the light of different evolutionary pressures and adaptive traits; 3) bursts of gene duplication events in specific lineages, like the cichlids (Brawand et al., 2014) has been suggested to have contributed to evolutionary novelty of the EGFR system in fish species. 4) More than 80 fish genomes have been recently deposited in online databases.

My analysis of fish EGFR veered on finding patterns of protein sequence evolution that could point to functional transitions. In the following paper, I identified a swapping pattern between opposite paralogs for a particular subclade of teleost fish, the cypriniformes. In their two copies of EGFR, EGFRa and EGFRb, some positions are conserved as in the opposite paralog, compared to the rest of the phylogeny. I named this event "Inter-paralog inversion", and showed that it might be related to a swap of functions between the paralogs.

2.2. Published Article

Pascarelli S, Laurino P. Inter-paralog amino acid inversion events in large phylogenies of duplicated proteins. PLoS computational biology. 2022;18(4):e1010016.

2.3. Conclusions

While biological databases get increasingly bigger, our ability to make use of all this data has a slower pace. With the analysis performed in this chapter, I aimed at providing a method to follow protein function in wide phylogenies, in the context of gene duplication. The mechanism behind functional transitions between paralogs is still not clear. However, the inter-paralog inversions appear to be a possible signal of when it happens. Detecting these functional transitions has an impact in inference by homology and functional residue prediction.

3. Chapter 3

Molecular Dynamics Provides a Mechanistical Explanation of Protein Function

(Published in *Proceedings of the National Academy of Sciences* as *"Insertions and deletions mediated functional divergence of Rossmann fold enzymes"*, and in *Proteins Science* as *"Protein dynamics induced by cryptic genetic variations shape the fitness of alanine:glyoxylate aminotransferase"*)

3.1. Introduction

Previous studies showed how proteins are not rigid folds tightly linked to one structure and one function (James & Tawfik, 2003). Rather, protein dynamics is a fundamental property that directly contributes to promiscuity of function and, therefore, to protein evolvability (Meier & Özbek, 2007; Tokuriki & Tawfik, 2009). Protein dynamics can be effectively simulated using Molecular Dynamics (MD) for a variety of purposes (Martin Karplus & Petsko, 1990). In the works reported in here, I use standard MD to track protein function, first on a cofactor-binding Rossmann fold, and then on the medically relevant human Alanine::glyoxylate aminotransferase.

3.2. Published Article 1

Toledo-Patiño S, Pascarelli S, Uechi GI, Laurino P. Insertions and deletions mediated functional divergence of Rossmann fold enzymes. Proc Natl Acad Sci U S A. 2022;119(48):e2207965119. Epub 2022/11/24. doi: 10.1073/pnas.2207965119. PubMed PMID: 36417431.

3.3. Published Article 2

Dindo M, Pascarelli S, Chiasserini D, Grottelli S, Costantini C, Uechi GI, Giardina G, Laurino P, Cellini B. Structural dynamics shape the fitness window of alanine: glyoxylate aminotransferase. Protein Sci. 2022;31(5):e4303.

3.4. Conclusions

In this chapter, I reported the results of MD analysis of scDH and AGT proteins and their mutants. The simulations revealed that scDH cofactor specificity can be re-engineered by InDels in the the β1-loop-α1 region of the Rossmann fold. This MD substantiated a putative path of functional transition in Rossmann folds, which shows how the plasticity of

this folds might have contributed to their evolutionary success. On the other hand, the MD simulations of AGT minor allele showed crucial differences to the major allele, explaining why the minor allele is more susceptible to disease-causing variants. Overall, MD simulations are a promising method to understand protein functions in a structure-full scientific world of the future.

Conclusions

There are multiple ways to look at protein function, each one reflecting a different aspect of the knowledge of the system. As such, the aspect of function under examination determines the challenge of the problem and how effective a bioinformatic analysis will be. In this thesis, I focused on the causal interpretation of functionality mediated by physical contacts, direct or indirect, defined in the Pittsburgh model as the "Interactions" hierarchical order (Keeling, Garza, Nartey, & Carvunis, 2019). The need of a defined set of rules explaining protein function arose from the field of de novo gene emergence, where the possibility of a transition between non-functional and functional locus is studied. How function in defined is then crucial to calculate the feasibility and the probability of a de novo gene emergence event. In this thesis, I do not examine such events. Although, using one of the Pittsburgh model definitions will help to clarify the meaning of protein function and its evolutionary implications in this thesis. As such, the Pittsburgh model might prove to be useful for the entire protein sciences community.

In this work, I showed how an in-silico analysis of protein sequence, structure, and dynamics can be used to test the target protein's ability to interact with other proteins or small ligands, thereby predicting protein function. Initially, I developed a computational method to identify the specific functional residues that are involved in a function that is altered among paralogs (chapter one). The method makes use of the ortholog conjecture: the assumption that functions are more conserved among orthologs. The test case of this method was performed on the Epidermal Growth Factor Receptor (EGFR) and its ligands. The ligands of EGFR all share the ability to bind the receptor while inducing different pathway activations. The analysis of conservation among orthologs and paralogs and the coevolution measures highlighted some residues in the ligands that, when mutated, altered the signaling cascade induced by the binding. Successive validation showed that the binding to the receptor, the property shared by all paralogous ligands, was not affected. Instead, a difference was observed in the dynamics of the ligand-receptor complex, specifically at the dimerization arm. I hereby obtained relevant data about the binding modes and signal transduction of the EGFR-ligands system. However, the method developed in this chapter has limited applicability outside those cases where orthologs do not share the same function. Also, it might overlook enabling mutations that act at a distance from the active site, usually missed by the commonly used covariation methods (Ding et al., 2022).

In the next chapter, I focused on the case where the ortholog conjecture does not hold true. I constructed a model to describe protein evolution after gene duplication and I used it to study the duplication of the EGF receptor in the fish lineage. Through the application of the model, it was possible to observe a specific pattern in the duplicated protein sequence conservation, named "inter-paralog residue inversion", in a subgroup of fish. I hypothesize that this conservation pattern is a proxy of a functional swap between paralogous proteins, signaling a break between the ancestry (or orthology) and the functional relationship. By using this model, I showed a way to follow the complex functional relationships in families of paralogous proteins that underwent sub-functionalization. Additionally, the analysis I performed pointed to which residues might be responsible for the functional swap, and which sub-function (or activated pathway) is affected by the event.

Lastly, in chapter three, I presented how molecular dynamics provides an opportunity to go beyond sequence and structure analysis. In the NAD(P)/FAD-binding Rossmann folds, previous studies showed that a glycine-rich motif in the β1-loop-α1 region was required to form a network of hydrogen bonds responsible for stabilizing the cofactor binding loop (Dym & Eisenberg, 2001). By performing MD simulations of a selected dehydrogenase (scDH), I observed that only a subset of the hydrogen bonds was required for NAD binding.

In contrast to a previous report on the hydrogen bond pattern within the GxGxxG motif or NAD binding Rossmann folds (Kleiger & Eisenberg, 2002), we did not observe the bond between the first and the last glycine of the motif. Meanwhile, in the deletion mutant ΔscDH simulations, the missing interactions did not affect the SAM binding property, showing a possible way in which InDels could shape cofactor specificity of Rossmann fold enzymes. In the second paper of chapter three, I analyzed the protein dynamics of two variants of human alanine glyoxylate aminotransferase. The minor allele (AGT-Mi) differs by just two substitutions, P11L and I340M, to the major allele (AGT-Ma), distally from the active site. However, AGT-Mi was shown to have a reduced activity and is a susceptibility factor to disease-causing mutations for primary hyperoxaluria. My analysis showed how these two distal residues have a propagating effect on the protein dynamics that influence the active site and two helices lying at the edge of disorder. Overall, MD simulations provided a mechanistic explanation for the alterations of protein function, even when the evolutionary context was not sufficient.

In conclusion, my work shows three approaches that use protein sequence, structure, and dynamics to study protein functional divergence. The methods developed in the first two chapters to identify functional residues will become an added value for protein scientists. Though, further work will be required to test the generality of these methods. In particular, both DIRpred and DIRphy will need to be tested on more proteins, possibly from different species, and, for the latter, on more duplication events with a different time frame compared to the teleost specific whole genome duplication. Meanwhile, the approach used in the two works of chapter three is a demonstration of the applicability of Alphafold 2 model structures for the analysis of protein mutations using molecular dynamics. In this way, I obtained molecular insights that explained the experimental data observed for the two systems of interest. However, this approach could not be used in all those cases where one structural model could not be obtained or is not representative of the entire ensemble of structural variation of one protein, as could be for an intrinsically disordered protein or one with large structural rearrangements. In those cases, new methodologies will need to be developed.
Overall, my contributions to protein bioinformatics demonstrate that this is a thriving field that let us understand and foresee functions in the molecular world. Further developments will be necessary to harness the vast extension of biological data, but in exchange it might reveal deep truths about the great mechanism that generated complexity on Earth: evolution.

References

Alberts, B., Johnson, A., Wilson, J., Lewis, J., Hunt, T., Roberts, K., . . . Walter, P. (2008). *Molecular Biology of the Cell*: Garland Science.

Altenhoff, A. M., Garrayo-Ventas, J., Cosentino, S., Emms, D., Glover, N. M., Hernández-Plaza, A., . . . Dessimoz, C. (2020). The Quest for Orthologs benchmark service and consensus calls in 2020. *Nucleic Acids Research, 48*(W1), W538-W545. doi:10.1093/nar/gkaa308

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *J Mol Biol, 215*(3), 403-410. doi:10.1016/s0022-2836(05)80360-2

Alvarez-Ponce, D., Feyertag, F., & Chakraborty, S. (2017). Position Matters: Network Centrality Considerably Impacts Rates of Protein Evolution in the Human Protein-Protein Interaction Network. *Genome Biol Evol, 9*(6), 1742-1756. doi:10.1093/gbe/evx117
10.1093/gbe/evx117.

Anfinsen, C. B. (1973). Principles that Govern the Folding of Protein Chains. *Science, 181*(4096), 223-230. doi:10.1126/science.181.4096.223

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., . . . Sherlock, G. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet, 25*(1), 25-29. doi:10.1038/75556
10.1038/75556.

Ashkenazy, H., Abadi, S., Martz, E., Chay, O., Mayrose, I., Pupko, T., & Ben-Tal, N. (2016). ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules. *Nucleic Acids Research, 44*(W1), W344-W350. doi:10.1093/nar/gkw408

Ashkenazy, H., Erez, E., Martz, E., Pupko, T., & Ben-Tal, N. (2010). ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. *Nucleic Acids Res, 38*(Web Server issue), W529-533. doi:10.1093/nar/gkq399

Atkinson, H. J., Morris, J. H., Ferrin, T. E., & Babbitt, P. C. (2009). Using sequence similarity networks for visualization of relationships across diverse protein superfamilies. *PLoS One, 4*(2), e4345. doi:10.1371/journal.pone.0004345
10.1371/journal.pone.0004345. Epub 2009 Feb 3.

Baptista, A. M., Jonson, P. H., Hough, E., & Petersen, S. B. (1998). The origin of trypsin: evidence for multiple gene duplications in trypsins. *J Mol Evol, 47*(3), 353-362. doi:10.1007/pl00006393
10.1007/pl00006393.

Barberán, S., Martín-Durán, J. M., & Cebrià, F. (2016). Evolution of the EGFR pathway in Metazoa and its diversification in the planarian Schmidtea mediterranea. *Sci Rep, 6*, 28071. doi:10.1038/srep28071

Barrett, A. J. (1997). Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (NC-IUBMB). Enzyme Nomenclature. Recommendations 1992. Supplement 4: corrections and additions (1997). *Eur J Biochem, 250*(1), 1-6. doi:10.1111/j.1432-1033.1997.001_1.x

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., . . . Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Res, 28*(1), 235-242. doi:10.1093/nar/28.1.235

Bernardi, R. C., Melo, M. C. R., & Schulten, K. (2015). Enhanced sampling techniques in molecular dynamics simulations of biological systems. *Biochim Biophys Acta, 1850*(5), 872-877. doi:10.1016/j.bbagen.2014.10.019

Bjorkelund, H., Gedda, L., & Andersson, K. (2011). Comparing the epidermal growth factor interaction with four different cell lines: intriguing effects imply strong dependency of cellular context. *PLoS One, 6*(1), e16536. doi:10.1371/journal.pone.0016536

Björklund, Å. K., Ekman, D., Light, S., Frey-Skött, J., & Elofsson, A. (2005). Domain Rearrangements in Protein Evolution. *Journal of Molecular Biology, 353*(4), 911-923. doi:https://doi.org/10.1016/j.jmb.2005.08.067

Bloom, J. D., & Arnold, F. H. (2009). In the light of directed evolution: Pathways of adaptive protein evolution. *Proceedings of the National Academy of Sciences, 106*(supplement_1), 9995-10000. doi:10.1073/pnas.0901522106

Brawand, D., Wagner, C. E., Li, Y. I., Malinsky, M., Keller, I., Fan, S., . . . Di Palma, F. (2014). The genomic substrate for adaptive radiation in African cichlid fish. *Nature, 513*, 375. doi:10.1038/nature13726 https://www.nature.com/articles/nature13726#supplementary-information

Brooks, B. R., Brooks, C. L., 3rd, Mackerell, A. D., Jr., Nilsson, L., Petrella, R. J., Roux, B., . . . Karplus, M. (2009). CHARMM: the biomolecular simulation program. *J Comput Chem, 30*(10), 1545-1614. doi:10.1002/jcc.21287

Brunet, F. G., Volff, J. N., & Schartl, M. (2016). Whole Genome Duplications Shaped the Receptor Tyrosine Kinase Repertoire of Jawed Vertebrates. In *Genome Biol Evol* (Vol. 8, pp. 1600-1613).

Caffrey, D. R., Somaroo, S., Hughes, J. D., Mintseris, J., & Huang, E. S. (2004). Are protein-protein interfaces more conserved in sequence than the rest of the protein surface? *Protein Sci, 13*(1), 190-202. doi:10.1110/ps.03323604

Capra, J. A., & Singh, M. (2007). Predicting functionally important residues from sequence conservation. *Bioinformatics, 23*(15), 1875-1882. doi:10.1093/bioinformatics/btm270

Chakrabarti, S., Bryant, S. H., & Panchenko, A. R. (2007). Functional specificity lies within the properties and evolutionary changes of amino acids. *J Mol Biol, 373*(3), 801-810. doi:10.1016/j.jmb.2007.08.036 10.1016/j.jmb.2007.08.036. Epub 2007 Aug 22.

Chakrabarti, S., & Panchenko, A. R. (2009). Ensemble approach to predict specificity determinants: benchmarking and validation. *BMC Bioinformatics, 10*, 207. doi:10.1186/1471-2105-10-207 10.1186/1471-2105-10-207.

Chakraborty, A., & Chakrabarti, S. (2014). A survey on prediction of specificity-determining sites in proteins. *Briefings in Bioinformatics, 16*(1), 71-88. doi:10.1093/bib/bbt092

Chen, X., & Zhang, J. (2012). The Ortholog Conjecture Is Untestable by the Current Gene Ontology but Is Supported by RNA Sequencing Data. *PLOS Computational Biology, 8*(11), e1002784. doi:10.1371/journal.pcbi.1002784

Copley, S. D. (2012). Moonlighting is mainstream: paradigm adjustment required. *Bioessays, 34*(7), 578-588. doi:10.1002/bies.201100191 10.1002/bies.201100191.

Dayhoff M, O. (1972). A model of evolutionary change in proteins. *Atlas of Protein Sequence and Structure, 5*, 89-99. Retrieved from https://cir.nii.ac.jp/crid/1570572700735867392

De Juan, D., Pazos, F., & Valencia, A. (2013). Emerging methods in protein co-evolution. *Nature Reviews Genetics, 14*(4), 249-261. doi:10.1038/nrg3414

De Vivo, M., Masetti, M., Bottegoni, G., & Cavalli, A. (2016). Role of Molecular Dynamics and Related Methods in Drug Discovery. *Journal of Medicinal Chemistry, 59*(9), 4035-4061. doi:10.1021/acs.jmedchem.5b01684

Dindo, M., Pascarelli, S., Chiasserini, D., Grottelli, S., Costantini, C., Uechi, G. I., . . . Cellini, B. (2022). Structural dynamics shape the fitness window of alanine: glyoxylate aminotransferase. *Protein Science, 31*(5), e4303.

Ding, D., Green, A. G., Wang, B., Lite, T.-L. V., Weinstein, E. N., Marks, D. S., & Laub, M. T. (2022). Co-evolution of interacting proteins through non-contacting and non-specific mutations. *Nature Ecology & Evolution, 6*(5), 590-603. doi:10.1038/s41559-022-01688-0

Dokholyan, N. V., Shakhnovich, B., & Shakhnovich, E. I. (2002). Expanding protein universe and its origin from the biological Big Bang. *Proceedings of the National Academy of Sciences, 99*(22), 14132-14136. doi:doi:10.1073/pnas.202497999

Dunker, A. K., Lawson, J. D., Brown, C. J., Williams, R. M., Romero, P., Oh, J. S., . . . Obradovic, Z. (2001). Intrinsically disordered protein. *J Mol Graph Model, 19*(1), 26-59. doi:10.1016/s1093-3263(00)00138-8

Durbin, R., Richard, D., Eddy, S. R., Eddy, S., Eddy, R., Krogh, A., & Mitchison, G. (1998). *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*: Cambridge University Press.

Dym, O., & Eisenberg, D. (2001). Sequence-structure analysis of FAD-containing proteins. *Protein Science, 10*(9), 1712-1728.

Eck, R. V., & Dayhoff, M. O. (1966). Evolution of the structure of ferredoxin based on living relics of primitive amino Acid sequences. *Science, 152*(3720), 363-366. doi:10.1126/science.152.3720.363

Eddy, S. R., Mitchison, G., & Durbin, R. (1995). Maximum Discrimination Hidden Markov Models of Sequence Consensus. *Journal of Computational Biology, 2*(1), 9-23. doi:10.1089/cmb.1995.2.9

Finn, R. D., Tate, J., Mistry, J., Coggill, P. C., Sammut, S. J., Hotz, H. R., . . . Bateman, A. (2008). The Pfam protein families database. *Nucleic Acids Res, 36*(Database issue), D281-288. doi:10.1093/nar/gkm960

Fischer, J., Fernández Ortuño, E., Marsoner, F., Artioli, A., Peters, J., Namba, T., . . . Heide, M. (2022). Human-specific ARHGAP11B ensures human-like basal progenitor levels in hominid cerebral organoids. *EMBO Rep, 23*(11), e54728. doi:10.15252/embr.202254728

Fitch, W. M. (1970). Distinguishing Homologous from Analogous Proteins. *Systematic Biology, 19*(2), 99-113. doi:10.2307/2412448

Force, A., Lynch, M., Pickett, F. B., Amores, A., Yan, Y. L., & Postlethwait, J. (1999). Preservation of duplicate genes by complementary, degenerative mutations. *Genetics, 151*(4), 1531-1545. Retrieved from http://dx.doi.org/

Gabaldón, T., & Koonin, E. V. (2013). Functional and evolutionary implications of gene orthology. *Nature Reviews Genetics, 14*(5), 360-366. doi:10.1038/nrg3456

Guo, H. H., Choe, J., & Loeb, L. A. (2004). Protein tolerance to random amino acid change. *Proceedings of the National Academy of Sciences, 101*(25), 9205-9210. doi:10.1073/pnas.0403255101

Harpole, T. J., & Delemotte, L. (2018). Conformational landscapes of membrane proteins delineated by enhanced sampling molecular dynamics simulations. *Biochim Biophys Acta Biomembr, 1860*(4), 909-926. doi:10.1016/j.bbamem.2017.10.033

Hase, T., Tanaka, H., Suzuki, Y., Nakagawa, S., & Kitano, H. (2009). Structure of Protein Interaction Networks and Their Implications on Drug Design. *PLOS Computational Biology, 5*(10), e1000550. doi:10.1371/journal.pcbi.1000550

Henikoff, S., & Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A, 89*(22), 10915-10919. doi:10.1073/pnas.89.22.10915

Hensen, U., Meyer, T., Haas, J., Rex, R., Vriend, G., & Grubmüller, H. (2012). Exploring protein dynamics space: the dynasome as the missing link between protein structure and function. *PLoS One, 7*(5), e33931. doi:10.1371/journal.pone.0033931

Higgins, D. G., Thompson, J. D., & Gibson, T. J. (1996). Using CLUSTAL for multiple sequence alignments. *Methods Enzymol, 266*, 383-402. doi:10.1016/s0076-6879(96)66024-8

Hollingsworth, S. A., & Dror, R. O. (2018). Molecular Dynamics Simulation for All. *Neuron, 99*(6), 1129-1143. doi:10.1016/j.neuron.2018.08.011

Hotaling, S., Kelley, J. L., & Frandsen, P. B. (2021). Toward a genome sequence for every animal: Where are we now? *Proceedings of the National Academy of Sciences, 118*(52), e2109019118. doi:10.1073/pnas.2109019118

Hughes, A. L., & Friedman, R. (2005). Gene duplication and the properties of biological networks. *J Mol Evol, 61*(6), 758-764. doi:10.1007/s00239-005-0037-z

Jacobs, T. M., Williams, B., Williams, T., Xu, X., Eletsky, A., Federizon, J. F., . . . Kuhlman, B. (2016). Design of structurally distinct proteins using strategies inspired by evolution. *Science, 352*(6286), 687-690. doi:10.1126/science.aad8036

James, L. C., & Tawfik, D. S. (2003). Conformational diversity and protein evolution – a 60-year-old hypothesis revisited. *Trends in Biochemical Sciences, 28*(7), 361-368. doi:https://doi.org/10.1016/S0968-0004(03)00135-X

Jeffery, C. J. (1999). Moonlighting proteins. *Trends in Biochemical Sciences, 24*(1), 8-11. doi:https://doi.org/10.1016/S0968-0004(98)01335-8

Jeffery, C. J. (2003). Moonlighting proteins: old proteins learning new tricks. *Trends Genet, 19*(8), 415-417. doi:10.1016/s0168-9525(03)00167-7 10.1016/S0168-9525(03)00167-7.

Johnson, L. K., Baxter, J. D., Vlodavsky, I., & Gospodarowicz, D. (1980). Epidermal growth factor and expression of specific genes: effects on cultured rat pituitary cells are dissociable from the mitogenic response. *Proc Natl Acad Sci U S A, 77*(1), 394-398. doi:10.1073/pnas.77.1.394 10.1073/pnas.77.1.394.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., . . . Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature, 596*(7873), 583-589. doi:10.1038/s41586-021-03819-2

Karplus, M., & Kuriyan, J. (2005). Molecular dynamics and protein function. *Proceedings of the National Academy of Sciences, 102*(19), 6679-6685. doi:10.1073/pnas.0408930102

Karplus, M., & McCammon, J. A. (2002). Molecular dynamics simulations of biomolecules. *Nat Struct Biol, 9*(9), 646-652. doi:10.1038/nsb0902-646

Karplus, M., & Petsko, G. A. (1990). Molecular dynamics simulations in biology. *Nature, 347*(6294), 631-639. doi:10.1038/347631a0

Keeling, D. M., Garza, P., Nartey, C. M., & Carvunis, A. R. (2019). The meanings of 'function' in biology and the problematic case of de novo gene emergence. *Elife, 8*. doi:10.7554/eLife.47014 10.7554/eLife.47014.

Kimura, M. (1991). The neutral theory of molecular evolution: a review of recent evidence. *Jpn J Genet, 66*(4), 367-386. doi:10.1266/jjg.66.367 10.1266/jjg.66.367.

Kleiger, G., & Eisenberg, D. (2002). GXXXG and GXXXA motifs stabilize FAD and NAD(P)-binding Rossmann folds through C(alpha)-H... O hydrogen bonds and van

der waals interactions. *J Mol Biol, 323*(1), 69-76. doi:10.1016/s0022-2836(02)00885-9

Koonin, E. V., Wolf, Y. I., & Karev, G. P. (2002). The structure of the protein universe and genome evolution. *Nature, 420*(6912), 218-223. doi:10.1038/nature01256

Korendovych, I. V. (2018). Rational and Semirational Protein Design. *Methods Mol Biol, 1685*, 15-23. doi:10.1007/978-1-4939-7366-8_2

Laisney, J., Braasch, I., Walter, R. B., Meierjohann, S., & Schartl, M. (2010). Lineage-specific co-evolution of the Egf receptor/ligand signaling system. *Bmc Evolutionary Biology, 10*, 16. doi:10.1186/1471-2148-10-27

Lane, J. R., May, L. T., Parton, R. G., Sexton, P. M., & Christopoulos, A. (2017). A kinetic view of GPCR allostery and biased agonism. *Nature Chemical Biology, 13*(9), 929-937. doi:10.1038/nchembio.2431

Lee, D., Redfern, O., & Orengo, C. (2007). Predicting protein function from sequence and structure. *Nat Rev Mol Cell Biol, 8*(12), 995-1005. doi:10.1038/nrm2281

Leman, J. K., Weitzner, B. D., Lewis, S. M., Adolf-Bryfogle, J., Alam, N., Alford, R. F., . . . Bonneau, R. (2020). Macromolecular modeling and design in Rosetta: recent methods and frameworks. *Nature Methods, 17*(7), 665-680. doi:10.1038/s41592-020-0848-2

Levin, K. B., Dym, O., Albeck, S., Magdassi, S., Keeble, A. H., Kleanthous, C., & Tawfik, D. S. (2009). Following evolutionary paths to protein-protein interactions with high affinity and selectivity. *Nature Structural & Molecular Biology, 16*(10), 1049-1055. doi:10.1038/nsmb.1670

Li, B., Fooksa, M., Heinze, S., & Meiler, J. (2018). Finding the needle in the haystack: towards solving the protein-folding problem computationally. *Crit Rev Biochem Mol Biol, 53*(1), 1-28. doi:10.1080/10409238.2017.1380596 10.1080/10409238.2017.1380596. Epub 2017 Oct 4.

Lichtarge, O., Bourne, H. R., & Cohen, F. E. (1996). An Evolutionary Trace Method Defines Binding Surfaces Common to Protein Families. *Journal of Molecular Biology, 257*(2), 342-358. doi:https://doi.org/10.1006/jmbi.1996.0167

Loewenstein, Y., Raimondo, D., Redfern, O. C., Watson, J., Frishman, D., Linial, M., . . . Tramontano, A. (2009). Protein function annotation by homology-based inference. *Genome biology, 10*(2), 207. doi:10.1186/gb-2009-10-2-207

Marchler-Bauer, A., Anderson, J. B., Derbyshire, M. K., DeWeese-Scott, C., Gonzales, N. R., Gwadz, M., . . . Bryant, S. H. (2007). CDD: a conserved domain database for interactive domain family analysis. *Nucleic Acids Res, 35*(Database issue), D237-240. doi:10.1093/nar/gkl951

Marrink, S. J., & Tieleman, D. P. (2013). Perspective on the Martini model. *Chem Soc Rev, 42*(16), 6801-6822. doi:10.1039/c3cs60093a

Mayrose, I., Graur, D., Ben-Tal, N., & Pupko, T. (2004). Comparison of site-specific rate-inference methods for protein sequences: empirical Bayesian methods are superior. *Mol Biol Evol, 21*(9), 1781-1791. doi:10.1093/molbev/msh194

Mazal, H., & Haran, G. (2019). Single-molecule FRET methods to study the dynamics of proteins at work. *Curr Opin Biomed Eng, 12*, 8-17. doi:10.1016/j.cobme.2019.08.007

McLachlan, A. D. (1979). Gene duplications in the structural evolution of chymotrypsin. *J Mol Biol, 128*(1), 49-79. doi:10.1016/0022-2836(79)90308-5 10.1016/0022-2836(79)90308-5.

Meier, S., & Özbek, S. (2007). A biological cosmos of parallel universes: Does protein structural plasticity facilitate evolution? *Bioessays, 29*(11), 1095-1104. doi:https://doi.org/10.1002/bies.20661

Meiler, J., & Baker, D. (2006). ROSETTALIGAND: Protein–small molecule docking with full side-chain flexibility. *Proteins: Structure, Function, and Bioinformatics, 65*(3), 538-548. doi:https://doi.org/10.1002/prot.21086

Moore, A. D., Björklund, Å. K., Ekman, D., Bornberg-Bauer, E., & Elofsson, A. (2008). Arrangements in the modular evolution of proteins. *Trends in Biochemical Sciences, 33*(9), 444-451. doi:https://doi.org/10.1016/j.tibs.2008.05.008

Morcos, F., Pagnani, A., Lunt, B., Bertolino, A., Marks, D. S., Sander, C., . . . Weigt, M. (2011). Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proceedings of the National Academy of Sciences, 108*(49), E1293-E1301. doi:10.1073/pnas.1111471108

Needleman, S. B., & Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol, 48*(3), 443-453. doi:10.1016/0022-2836(70)90057-4

Nehrt, N. L., Clark, W. T., Radivojac, P., & Hahn, M. W. (2011). Testing the Ortholog Conjecture with Comparative Functional Genomic Data from Mammals. *PLOS Computational Biology, 7*(6), e1002073. doi:10.1371/journal.pcbi.1002073

Nemoto, W., Saito, A., & Oikawa, H. (2013). Recent advances in functional region prediction by using structural and evolutionary information - Remaining problems and future extensions. *Comput Struct Biotechnol J, 8*, e201308007. doi:10.5936/csbj.201308007

Nowell, P. C., & Hungerford, D. A. (1960). Chromosome studies on normal and leukemic human leukocytes. *J Natl Cancer Inst, 25*, 85-109.

Oeller, M., Kang, R., Sormanni, P., & Vendruscolo, M. (2022). Sequence-based pH-dependent prediction of protein solubility using CamSol. *bioRxiv*, 2022.2005.2009.491135. doi:10.1101/2022.05.09.491135

Ohno, S. (1970). *Evolution by gene duplication*: Springer-Verlag.

Oldfield, C. J., & Dunker, A. K. (2014). Intrinsically disordered proteins and intrinsically disordered protein regions. *Annu Rev Biochem, 83*, 553-584. doi:10.1146/annurev-biochem-072711-164947

Pascarelli, S., & Laurino, P. (2022). Inter-paralog amino acid inversion events in large phylogenies of duplicated proteins. *PLOS Computational Biology, 18*(4), e1010016.

Pascarelli, S., Merzhakupova, D., Uechi, G.-I., & Laurino, P. (2020). Single EGF mutants unravel the mechanism for stabilization of Epidermal Growth Factor Receptor (EGFR) system. *bioRxiv*, 677393. doi:10.1101/677393

Patil, S. S., Catanese, H. N., Brayton, K. A., Lofgren, E. T., & Gebremedhin, A. H. (2022). Sequence Similarity Network Analysis Provides Insight into the Temporal and Geographical Distribution of Mutations in SARS-CoV-2 Spike Protein. *Viruses, 14*(8). doi:10.3390/v14081672

Perdigão, N., Heinrich, J., Stolte, C., Sabir, K. S., Buckley, M. J., Tabor, B., . . . O'Donoghue, S. I. (2015). Unexpected features of the dark proteome. *Proceedings of the National Academy of Sciences, 112*(52), 15898-15903. doi:10.1073/pnas.1508380112

Pereira, J., Simpkin, A. J., Hartmann, M. D., Rigden, D. J., Keegan, R. M., & Lupas, A. N. (2021). High-accuracy protein structure prediction in CASP14. *Proteins: Structure, Function, and Bioinformatics, 89*(12), 1687-1699. doi:https://doi.org/10.1002/prot.26171

Piatigorsky, J. (2007). *Gene Sharing and Evolution*: Harvard University Press.

Pongsupasa, V., Anuwan, P., Maenpuen, S., & Wongnate, T. (2022). Rational-Design Engineering to Improve Enzyme Thermostability. *Methods Mol Biol, 2397*, 159-178. doi:10.1007/978-1-0716-1826-4_9

Pupko, T., Bell, R. E., Mayrose, I., Glaser, F., & Ben-Tal, N. (2002). Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. *Bioinformatics, 18 Suppl 1*, S71-77. doi:10.1093/bioinformatics/18.suppl_1.s71 10.1093/bioinformatics/18.suppl_1.s71.

Rastogi, S., & Liberles, D. A. (2005). Subfunctionalization of duplicated genes as a transition state to neofunctionalization. *BMC Evol Biol, 5*, 28. doi:10.1186/1471-2148-5-28

Remmert, M., Biegert, A., Hauser, A., & Söding, J. (2012). HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nature Methods, 9*(2), 173-175. doi:10.1038/nmeth.1818

Rojano, E., Jabato, F. M., Perkins, J. R., Córdoba-Caballero, J., García-Criado, F., Sillitoe, I., . . . Seoane-Zonjic, P. (2022). Assigning protein function from domain-function associations using DomFun. *BMC Bioinformatics, 23*(1), 43. doi:10.1186/s12859-022-04565-6

Rost, B. (1999). Twilight zone of protein sequence alignments. *Protein Engineering, Design and Selection, 12*(2), 85-94. doi:10.1093/protein/12.2.85

Roth, C., Rastogi, S., Arvestad, L., Dittmar, K., Light, S., Ekman, D., & Liberles, D. A. (2007). Evolution after gene duplication: models, mechanisms, sequences, systems, and organisms. *J Exp Zool B Mol Dev Evol, 308*(1), 58-73. doi:10.1002/jez.b.21124 10.1002/jez.b.21124.

Rueda, M., Ferrer-Costa, C., Meyer, T., Pérez, A., Camps, J., Hospital, A., . . . Orozco, M. (2007). A consensus view of protein dynamics. *Proceedings of the National Academy of Sciences, 104*(3), 796-801. doi:10.1073/pnas.0605534104

Schaeffer, R. D., & Daggett, V. (2011). Protein folds and protein folding. *Protein Eng Des Sel, 24*(1-2), 11-19. doi:10.1093/protein/gzq096

Senior, A. W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., . . . Hassabis, D. (2020). Improved protein structure prediction using potentials from deep learning. *Nature, 577*(7792), 706-710. doi:10.1038/s41586-019-1923-7 10.1038/s41586-019-1923-7. Epub 2020 Jan 15.

Senn, H. M., & Thiel, W. (2009). QM/MM methods for biomolecular systems. *Angew Chem Int Ed Engl, 48*(7), 1198-1229. doi:10.1002/anie.200802019

Sikosek, T., Chan, H. S., & Bornberg-Bauer, E. (2012). Escape from Adaptive Conflict follows from weak functional trade-offs and mutational robustness. *Proc Natl Acad Sci U S A, 109*(37), 14888-14893. doi:10.1073/pnas.1115620109

Skwark, M. J., Abdel-Rehim, A., & Elofsson, A. (2013). PconsC: combination of direct information methods and alignments improves contact prediction. *Bioinformatics, 29*(14), 1815-1816. doi:10.1093/bioinformatics/btt259

Song, Y., DiMaio, F., Wang, R. Y., Kim, D., Miles, C., Brunette, T., . . . Baker, D. (2013). High-resolution comparative modeling with RosettaCM. *Structure, 21*(10), 1735-1742. doi:10.1016/j.str.2013.08.005

Sonnhammer, E. L. L., & Hollich, V. (2005). Scoredist: A simple and robust protein sequence distance estimator. *BMC Bioinformatics, 6*(1), 108. doi:10.1186/1471-2105-6-108

Stamboulian, M., Guerrero, R. F., Hahn, M. W., & Radivojac, P. (2020). The ortholog conjecture revisited: the value of orthologs and paralogs in function prediction. *Bioinformatics, 36*(Supplement_1), i219-i226. doi:10.1093/bioinformatics/btaa468

Storz, J. F. (2016). Gene Duplication and Evolutionary Innovations in Hemoglobin-Oxygen Transport. In *Physiology (Bethesda)* (Vol. 31, pp. 223-232).

Sun, Z., Liu, Q., Qu, G., Feng, Y., & Reetz, M. T. (2019). Utility of B-Factors in Protein Science: Interpreting Rigidity, Flexibility, and Internal Motion and Engineering Thermostability. *Chem Rev, 119*(3), 1626-1665. doi:10.1021/acs.chemrev.8b00290

Suzek, B. E., Huang, H., McGarvey, P., Mazumder, R., & Wu, C. H. (2007). UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics, 23*(10), 1282-1288. doi:10.1093/bioinformatics/btm098

Thomas, P. D., Ebert, D., Muruganujan, A., Mushayahama, T., Albou, L. P., & Mi, H. (2022). PANTHER: Making genome-scale phylogenetics accessible to all. *Protein Sci, 31*(1), 8-22. doi:10.1002/pro.4218

Thompson, J. D., Higgins, D. G., & Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res, 22*(22), 4673-4680. doi:10.1093/nar/22.22.4673

Thompson, L. R., Sanders, J. G., McDonald, D., Amir, A., Ladau, J., Locey, K. J., . . . The Earth Microbiome Project, C. (2017). A communal catalogue reveals Earth's multiscale microbial diversity. *Nature, 551*(7681), 457-463. doi:10.1038/nature24621

Tokuriki, N., & Tawfik, D. S. (2009). Protein Dynamism and Evolvability. *Science, 324*(5924), 203-207. doi:10.1126/science.1169375

Toledo-Patiño, S., Pascarelli, S., Uechi, G. I., & Laurino, P. (2022). Insertions and deletions mediated functional divergence of Rossmann fold enzymes. *Proc Natl Acad Sci U S A, 119*(48), e2207965119. doi:10.1073/pnas.2207965119

Van Der Spoel, D., Lindahl, E., Hess, B., Groenhof, G., Mark, A. E., & Berendsen, H. J. (2005). GROMACS: fast, flexible, and free. *J Comput Chem, 26*(16), 1701-1718. doi:10.1002/jcc.20291

Volff, J. N. (2005). Genome evolution and biodiversity in teleost fish. *Heredity (Edinb), 94*(3), 280-294. doi:10.1038/sj.hdy.6800635 10.1038/sj.hdy.6800635.

Wang, J., Wolf, R. M., Caldwell, J. W., Kollman, P. A., & Case, D. A. (2004). Development and testing of a general amber force field. *J Comput Chem, 25*(9), 1157-1174. doi:10.1002/jcc.20035

Wang, Y., Zhang, H., Zhong, H., & Xue, Z. (2021). Protein domain identification methods and online resources. *Comput Struct Biotechnol J, 19*, 1145-1153. doi:https://doi.org/10.1016/j.csbj.2021.01.041

Wee, P., & Wang, Z. (2017). Epidermal Growth Factor Receptor Cell Proliferation Signaling Pathways. *Cancers (Basel), 9*(5). doi:10.3390/cancers9050052

Wilson, K. J., Gilmore, J. L., Foley, J., Lemmon, M. A., & Riese, D. J., 2nd. (2009). Functional selectivity of EGF family peptide growth factors: implications for cancer. *Pharmacol Ther, 122*(1), 1-8. doi:10.1016/j.pharmthera.2008.11.008 10.1016/j.pharmthera.2008.11.008. Epub 2008 Dec 16.

Wüthrich, K. (2001). The way to NMR structures of proteins. *Nat Struct Biol, 8*(11), 923-925. doi:10.1038/nsb1101-923

Yip, S. H. C., Foo, J.-L., Schenk, G., Gahan, L. R., Carr, P. D., & Ollis, D. L. (2011). Directed evolution combined with rational design increases activity of GpdQ toward a non-physiological substrate and alters the oligomeric structure of the enzyme. *Protein Engineering, Design and Selection, 24*(12), 861-872. doi:10.1093/protein/gzr048

Yu, L., Zhang, Y., Gutman, I., Shi, Y., & Dehmer, M. (2017). Protein Sequence Comparison Based on Physicochemical Properties and the Position-Feature Energy Matrix. *Scientific Reports, 7*(1), 46237. doi:10.1038/srep46237