# Identification of residue inversions in large phylogenies of duplicated proteins

Stefano Pascarelli[1], Paola Laurino[1*]

[1] Protein Engineering and Evolution Unit, Okinawa Institute of Science and Technology Graduate University, Onna, Okinawa 904-0495 Japan.

* Corresponding author

E-mail: paola.laurino@oist.jp

## Abstract

Connecting protein sequence to function is becoming increasingly relevant since high-throughput sequencing studies accumulate large amounts of genomic data. Protein database annotation helps to bridge this gap; however, it is fundamental to understand the mechanisms underlying functional inheritance and divergence. If the homology relationship between proteins is known, can we determine whether the function diverged?

In this work, we analyze different possibilities of protein sequence evolution after gene duplication and identify "residue inversions", i.e., sites where the relationship between the ancestry and the functional signal is decoupled. Residues in these sites play a role in functional divergence and could indicate a shift in protein function. We develop a method to recognize residue inversions in a phylogeny and test it on real and simulated datasets. In a dataset built from the Epidermal Growth Factor Receptor (EGFR) sequences found in 88 fish species, we identify 19 positions that went through inversion after gene duplication, mostly located at the ligand-binding extracellular domain.

## Introduction

Proteins perform their function either through protein-protein interactions, protein-ligand interactions, or catalyzing chemical reactions. At the molecular level, residues of a protein interact with a counterpart, namely a ligand (small molecule, protein, DNA/RNA, etc.). Herein we define such interacting residues as functional residues. The importance of predicting functional residues in a protein is evident as these residues can contribute to designing new functions, switching specificities, defining protein families and subfamilies, or identifying the occurrence of a functional innovation (e.g., a change of ligand specificity). Crystal structures in which the protein of interest was co-crystallized with its ligand, can readily identify functional residues. However, when the structure is not available, the identification of functional residues is not trivial. Previous attempts used the evolutionary information found in protein sequences and their homologs (Sankararaman, et al. 2009; Wilkins, et al. 2012; Ashkenazy, et al. 2016), approaches now facilitated by the global-scale genome sequencing effort driven by the development of high-throughput sequencing technologies. The prediction of functional residues by such methods is hampered by the presence of neutral mutations, namely amino acid substitutions that are neither beneficial nor disadvantageous (Kimura 1991). Non-synonymous neutral mutations are on average 10 times more abundant than advantageous mutations (Rockah-Shmuel, et al. 2015). While neutral mutations help to determine the phylogenetic position of a protein, mutations of functional residues are a signal of functional shifts that might occur independently of the phylogeny. When the relationship between phylogeny and function is decoupled, neutral mutations may mislead homology-based prediction methods, which are the most common way of functional prediction (Lee, et al. 2007).

An event that often decouples the phylogenetic and functional signal is gene duplication. After gene duplication, two proteins follow a semi-independent evolution. For example, before

diverging, the two duplicates may influence each other by gene conversion (Mitchell 1955) or homomeric-heteromeric interactions (Mallik and Tawfik 2020), and they tend to diversify in the expression profile (Huerta-Cepas, et al. 2011; Gout and Lynch 2015). Duplicated proteins are reported to go through functional diversification (Innan and Kondrashov 2010). For example, in an event termed "sub-functionalization", a protein with multiple functions (e.g., cellular receptors binding multiple ligands) might split its functions between the two gene copies after duplication. Previously, McClintock *et al.* showed that, in zebrafish HOX genes, the subset of functions inherited by the duplicated copies is different between fish and mouse – a phenomenon named "function shuffling" (McClintock, et al. 2002). In cases alike, the phylogenetic signal is misleading when used to predict the function of a "shuffled" orthologous protein. However, if the functional divergence is correctly identified, it allows to highlight the functional residues responsible for this transition, with reduced noise from the neutral variants. In this work, we address the identification of protein functional residues that are mutated during this type of functional rearrangement.

Residues responsible for a change of function within a protein family are usually called Specificity Determining Sites (SDS). SDS can be predicted by multiple methods (Chakraborty and Chakrabarti 2014). SDS prediction methods use the Multiple Sequence Alignment (MSA) or the 3D structure of the protein of interest to calculate a score based on conservation (Casari, et al. 1995; Lichtarge, et al. 1996; Hannenhalli and Russell 2000; Mirny and Gelfand 2002; Kalinina, et al. 2004; Panchenko, et al. 2004; Reva, et al. 2007; Fischer, et al. 2008; Wuster, et al. 2010), evolutionary rate (Abhiman and Sonnhammer 2005; Capra and Singh 2008; Gu, et al. 2013), or 3D structure properties (Amitai, et al. 2004). Most of these approaches require the user to provide the correct groupings of the homologous proteins. When this information is missing, the groupings

are made according to the ortholog classification obtained by manual or automatic partitioning methods (Wicker, et al. 2001; Storm and Sonnhammer 2002; Zmasek and Eddy 2002; Li, et al. 2003). However, the SDS predictions in automatically partitioned orthologs showed a lower sensitivity (Chakraborty, et al. 2012), demonstrating that an incorrect grouping negatively influences the prediction. The grouping usually follows the ortholog conjecture, namely that orthologs are more conserved than paralogs (Tatusov, et al. 1997; Dolinski and Botstein 2007). When this is not true, the SDS prediction is hampered. Therefore, the power to predict functional residues is limited by our ability to track protein function on the phylogenetic tree when it is not linearly inherited by orthologs. In our work we address this problem by identifying a signal of functional transition that might prove to be useful when annotating orthologs.

EGFR (Epidermal Growth Factor Receptor) is a tyrosine-kinase receptor that activates multiple signaling pathways after binding one of the seven EGFR ligands (Oda, et al. 2005). EGFR is broadly expressed (Yano, et al. 2003) and plays a crucial role in several aspects of organismal development and homeostasis like cellular growth, differentiation, metabolism, and motility (Hubbard and Miller 2007). In fish, two copies of EGFR were kept after the Teleost-Specific Genome Duplication (TSGD) event that occurred about 350 mya in the actinopterygian lineage (Amores, et al. 2004; Meyer and Van de Peer 2005). Lorin *et al.* showed that both copies of EGFR might have been retained because they are involved in the complex process of skin pigmentation (Lorin, et al. 2018), which is under high evolutionary pressure in most fish. Furthermore, the extracellular domain of fish EGFR, responsible for binding multiple ligands, likely went through sub-functionalization (Laisney, et al. 2010). For these reasons, EGFR constitutes a perfect model to study uneven functional inheritance events.

In this work, we observe a scenario where the function of a protein is not linearly inherited across orthologs, and we identify the functional residues responsible for the shift of functions. Our goal is to develop an algorithm that highlights the signature of a putative inversion of function, as could be an inversion of residues between the paralogs within the same species. First, we obtain a simple theoretical model that describes the likelihood of a residue inversion in comparison to other outcomes. Then, based on the model, we develop an algorithm that identifies residue inversions in a phylogeny, and we apply it in the context of fish EGFR duplication. Finally, we validate the results using statistical scores and simulated evolution. Our analysis shows a new way to investigate an important and understudied outcome of gene duplication.

## Results and Discussions

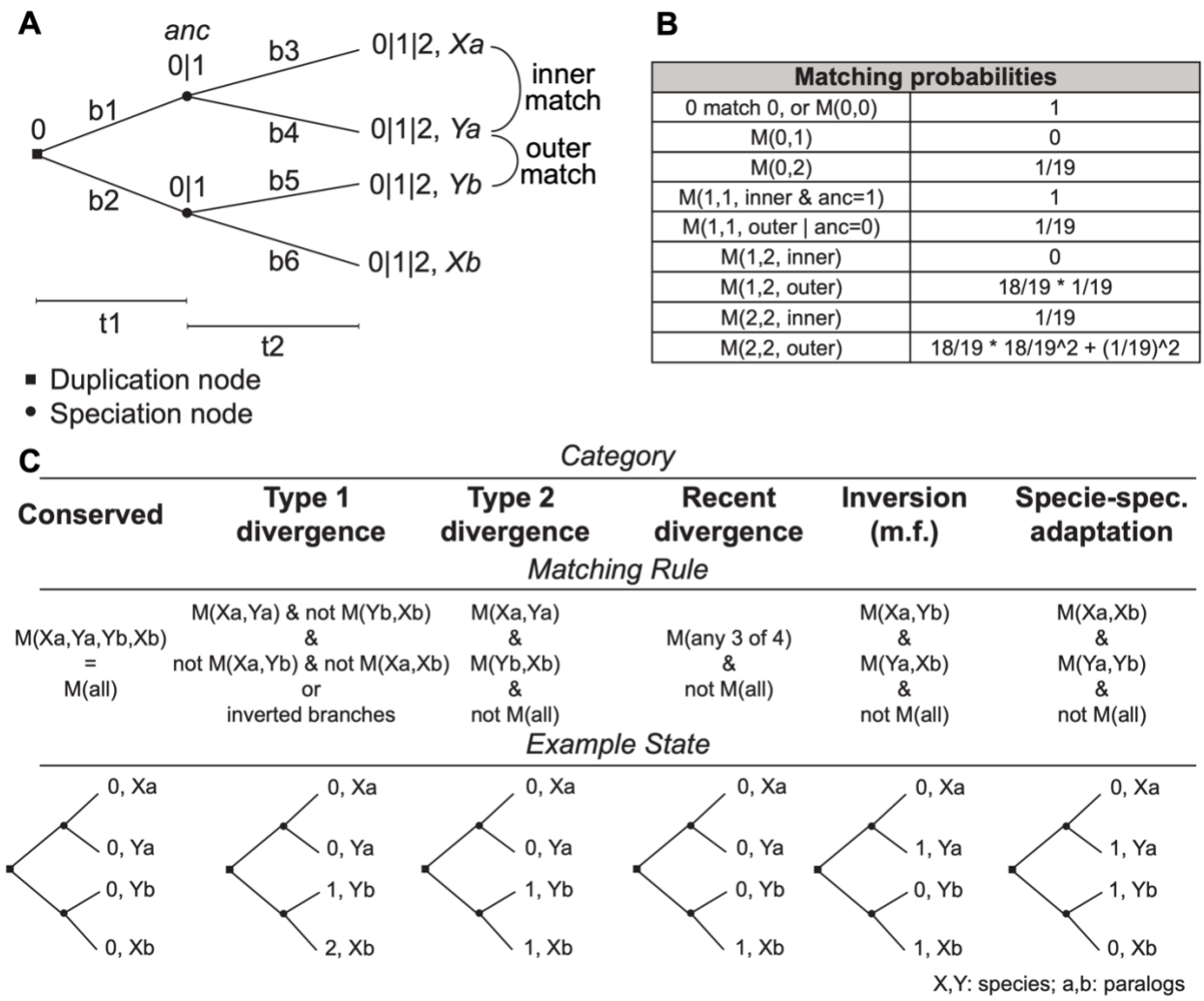
### Theoretical model for residue inversion identification

First, we constructed a simplified model to describe the evolution of a residue through a protein phylogeny after gene duplication and speciation. In the model, a tree branch (b1–b6) could be either mutation (1) or no-mutation (0) state, while a leaf node (Xa, Ya, Xb, Yb) has the possible states 0, 1, or 2 depending on the number of mutations in the preceding branches (Fig. 1A). A configuration of six branch states univocally leads to a configuration of four leaf nodes. The model uses two branch length parameters, pre-speciation (t1) and post-speciation (t2), to calculate the probability of each of the 64 branch configurations. Using the model and a set of matching rules between the leaves (Fig. 1B), we assessed the probability of seven categories/scenarios of configurations (Fig. 1C): *Conserved*, all four residues match; *Type 1* divergence, residues match

in only one paralog group; *Type 2* divergence, residues match per paralog group but not per species; *Recent divergence*, all residues match except for one leaf node; *Inversion*, residues of opposite paralog group match but not per species; *Species-specific adaptation*, residues match per species, but not per paralog group; and *Non-conserved*, collecting all events that do not fall in any of the other categories. We used the following formula to calculate the probability of each category:

$$P_{cat} = \sum_{s}^{N} P_{conf} * P_{(cat|conf)}$$

Where N are the 64 configurations of branch states, the probability $P_{conf}$ of the branch configuration is given by the model (Fig. S4), and the conditional probability for the category $P_{(cat|conf)}$ is determined by the category's matching rules.




| Matching probabilities | |
|---|---|
| 0 match 0, or M(0,0) | 1 |
| M(0,1) | 0 |
| M(0,2) | 1/19 |
| M(1,1, inner & anc=1) | 1 |
| M(1,1, outer \| anc=0) | 1/19 |
| M(1,2, inner) | 0 |
| M(1,2, outer) | 18/19 * 1/19 |
| M(2,2, inner) | 1/19 |
| M(2,2, outer) | 18/19 * 18/19^2 + (1/19)^2 |

Figure 1. Theoretical model of the evolution of protein residue after gene duplication. (A) The structure of the phylogenetic tree that the model is based on. The branch lengths t1 and t2 are used to determine the probability of a mutation on each branch b1 to b6. A leaf node can be found in states 0, 1, or 2 depending on the number of mutations in the preceding branches. An inner match is defined to be a match between orthologs (Xa to Ya, or Xb to Yb), while an outer match is any other match. The probability for a match between two states is given by the table in (B) and represents the underlying transition to any of the 20 amino acids. (C) Description of the categories. The categories represent a biologically interpretable situation, suggested by their name. Given a certain outcome configuration of states, it is possible to calculate the probability of observing a

certain category by using the matching rule. The “Example State” section shows the leaf configuration that gives the highest probability of observing the category described.

Next, we tested the probability of each category at varying branch lengths (Fig. 2A). As expected, short branch lengths lead to a high probability of conserved residues, and long branch lengths lead to non-conserved residues. When the pre-speciation branch length is the longest of the two, we observe a high probability of Type 1 and Type 2 divergence. Whereas, when the post-speciation branch length is the longest, we observe a high probability of recent divergence. Interestingly, for any branch length combination, the two categories of inversion and species-specific adaptation have less than a 1% chance of appearing. From this analysis, we expect to rarely find these two events in real data.

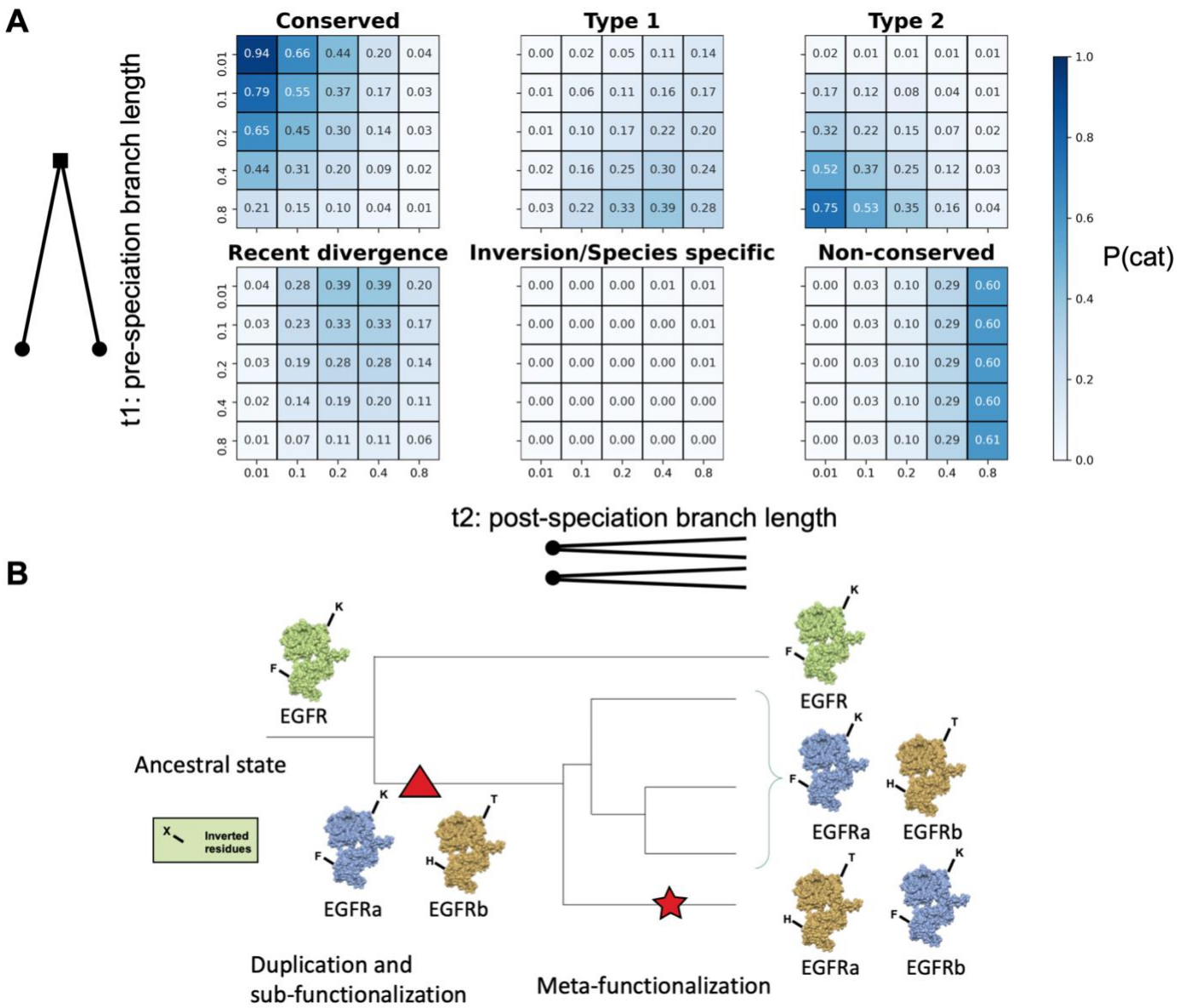



Figure 2. Theoretical model results. (A) The heatmaps show the category probabilities at different tree branch lengths as calculated using the theoretical model definitions. (B) Exemplification of "Meta-functionalization" (star), the putative driver of residue inversions in the phylogeny. A multifunctional protein (green) subdivides its functions (blue and yellow) between the two copies obtained after gene duplication (triangle). In a sub-group of species, the functional inheritance of the two copies is inverted. This event is revealed by the pattern of inverted residues compared to the majority of the other species.

From the previous results, residues inversions should not be recurrent in a phylogeny. Thus, an unlikely high presence of inverted residues may show that: 1) these residues are following a selective pressure directly related to the paralog function that, for a particular subgroup of species, is opposite and complementary to the other species; or 2) these residues act as compensatory mutations. A recurrent residue inversion could be the proxy of a functional rearrangement between the paralogs within a clade. A rearrangement of functions is facilitated in, but not limited to the Innovation Amplification Divergence (IAD) model of duplicates divergence (Bergthorsson, et al. 2007), where paralogs partially retain their secondary functions. We describe this subgroup-specific functional inversion event as a possible outcome after the sub-functionalization of a duplicated gene and we suggest the name "meta-functionalization" from the Greek word "metathesis", namely "put in a different order" (Fig. 2B).

**Algorithm for the identification of residue inversions**

We developed an algorithm to identify the events of inversion in a protein phylogeny using a multiple sequence alignment (MSA) and a phylogenetic tree. The algorithm was implemented as a python package named DIRphy, for the Detection of Inverted Residues in a phylogeny. DIRphy splits the protein sequences of the MSA into four groups according to the organism and orthology annotations, which can be either provided by the user or automatically done using a tree distance parameter. Based on the matching probabilities of the previous theoretical model, DIRphy calculates a score for each event of "Inversion" and "Species-specific adaptation", representing its probability to occur (see methods for details). However, for this paper, we will focus only on the inversion events. The output of DIRphy is a list of positions above a defined threshold. When the

organism grouping is manually selected, the script calculates both the observed and the expected probability of inversion between the specified groups in the given tree. Otherwise, when the organism grouping is not specified, the output table shows the observed probability of residue inversion given by the grouping that has the highest probability in that position. In the current version of DIRphy, only a binary paralog classification is allowed. DIRphy is released as an open-source project in Github: https://github.com/OISTpasca/protein-inversions

**Construction of a fish EGFR dataset and identification of residue inversions**

We tested DIRphy in the phylogenetic tree of the Epidermal Growth Factor Receptor (EGFR) in the fish lineage. First, we filtered 88 fish genomes (taxon 41665) from the European Nucleotide Archive (ENA) (Leinonen, et al. 2011) to obtain a dataset of 167 fish EGFR protein sequences. The dataset included all high-quality (N50 > 1Mb) teleost genomes plus one outgroup before the TSGD (spotted gar). From the phylogenetic analysis of this dataset (Supplementary figure 1), three clear duplication events can be observed. The first, most ancient duplication coincided with the TSGD, around the time of the split with gars (~350 mya), and resulted in the separation of the two copies: EGFRa and EGFRb. Two more copies were found in most salmonids, corresponding to the salmon-specific whole-genome duplication around ~80 mya (Macqueen and Johnston 2014), and one more copy in goldfish, possibly due to the carp-specific whole genome duplication about ~10 mya (Wang, et al. 2012). The longer branch lengths of EGFRb (Mann-Whitney test, p-val: 2.305e-11) indicate that EGFRb is evolving more rapidly than its counterpart. Furthermore, the EGFRb gene was more commonly lost. Out of 15 gene loss events, only one species lost EGFRa (*S grahami*), while 14 species lost EGFRb.

Next, we tested the previously computed phylogeny and MSA of fish EGFR for residue inversions using DIRphy. We decided to compare the Cypriniformes clade with the other teleosts because of the high coverage of genomes in both groups, and a sufficiently long separation between the two groups to allow functional divergence on the protein sequence (Fig. 3A). We observed a distribution of the scores that resembles an extreme-value distribution, with most of them below 0.01 (Fig. 3B). Using the theoretical model, we calculated that the expected value for the probability of residue inversions in the same tree is 0.002, much lower than the score of the 99th percentile of 0.16. This percentile value was used as a threshold to select eight sites, the majority of which (six) were from the EGFR extracellular domain. We highlighted the positions of the six sites on the 3D structure of EGFRa and EGFRb extracellular domain from a representative Cypriniformes (*S. anshuiensis*), modeled using the human structure template (Fig. 3C). Of the six selected positions, only one (MSA pos 520) was found at the ligand-binding pocket interface. Pos 520 corresponds to Phe-357 in hEGFR. Previous studies showed that the hydrophobic interaction between Phe-357 and Tyr-13 in the ligand hEGF is determinant for the binding (Tadaki and Niyogi 1993). In the fish clade *Xiphophorus*, the observed change between Phe and His at this position is considered to be the determinant cause of the different responses of EGFRa and EGFRb after ligand stimulation (Gomez, et al. 2004). Out of the six positions in the extra-cellular domain, two showed a conservative substitution (Fig. 3D). MSA position 506 contains hydrophobic and aliphatic amino acids (Ile or Val), while position 491 shows a small and uncharged amino acid (Ser or Thr). All other positions exhibited a shift of amino acid physicochemical properties. These results show that DIRphy can identify inverted residues in a protein duplication phylogeny, regardless of the amino acid substitution type.

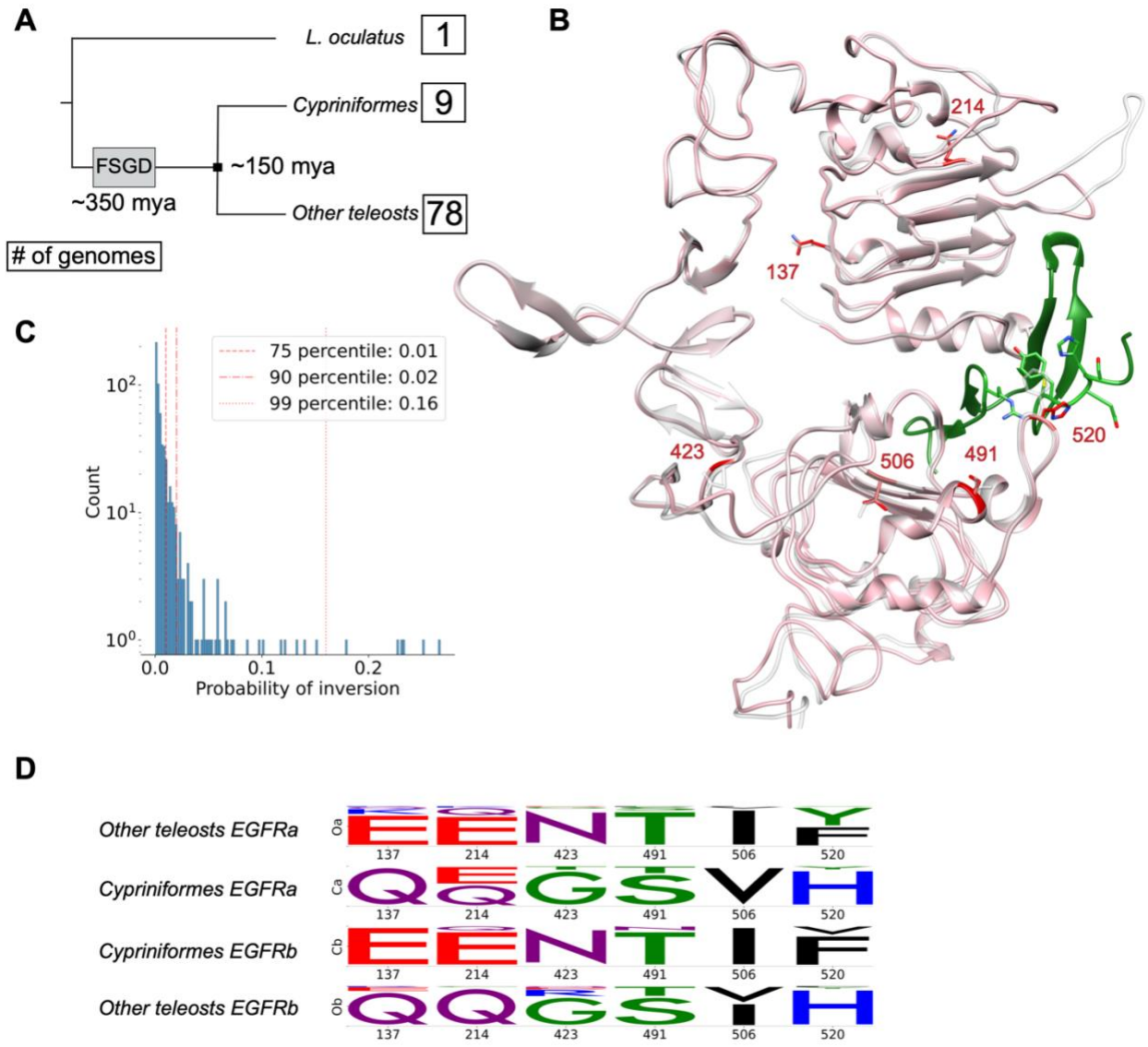



Fig. 3. Inverted residues in fish EGFR. (A) Schematic representation of the fish dataset phylogeny. The dates (mya: million years ago) indicate the time of the Fish Specific whole Genome Duplication (FSGD) and the separation of Cypriniformes fish to all other teleost fish. The number in the boxes represents how many genomes are in the dataset for that group. (B) 3D model superposition of *S anshuensis* EGFRa (pink) and EGFRb (white), generated by homology using human EGFR as a template (1IVO) (Ogiso, et al. 2002). The inverted residues have been

highlighted in red. The ligand EGF (green) was taken from the human model after superposing the receptors. (C) DIRphy score distribution. The residue inversion probability score was calculated for each site in the MSA that has less than 60% gaps. The top 1% of sites were further characterized. (D) Logo representation of the four sub alignments (two species groups, two protein copies) in the inverted residue sites. The logo represents the normalized amino acid count per column and was obtained using the python package Logomaker (Tareen and Kinney 2020).

**Score validation by simulated evolution**

We statistically validated the score observed in the fish EGFR data using a simulated evolution experiment. In this simulation, random starting amino acids are run through a phylogenetic tree that has the same topology as the previously computed fish EGFR tree. The simulation used the same evolutionary model of the fish EGFR tree to output a MSA as a result. Compared to the fish EGFR MSA, the simulated evolution MSA showed on average lower DIRphy scores while having a similar shape of the score distribution (Fig. 4A). No specific amino acid was found to have high scores. Interestingly, the three residues involved in the interaction in position 520 (His, Tyr, and, except for one site, Phe) failed to produce any score higher than 0.05 in the simulation (Fig. 4B). For further analysis, we used the 99$^{th}$ percentile score of the simulation as a threshold for selecting inverted residues in the real dataset (Supplementary table 1, Supplementary figure 2). In summary, the simulated evolution experiment provided a score threshold for detecting residue inversions and confirmed the low chance of this event in the fish EGFR dataset, as observed in the theoretical model.

**Tree bootstrap**

We characterized the phylogenetic information carried by the residue inversions when reconstructing the correct fish EGFR phylogeny tree structure. First, we computed a phylogenetic tree using the sub-alignment of 19 inverted residue sites that score higher than the 99$^{th}$ percentile in the simulated evolution experiment (Supplementary figure 3). In this phylogenetic tree, we can still see two defined groups of sequences corresponding to the two copies of EGFR. Though, as expected, all the sequences from Cypriniformes EGFRa cluster together with the EGFRb group, and vice versa the Cypriniformes EGFRb cluster in the EGFRa group. When we compared this tree to the full-alignment tree, we calculated a similarity value of 74.84 out of 165 using the tree similarity score of Nye *et al.* (Nye, et al. 2006). Next, we generated a pool of bootstrap trees with a reduced alignment length, and we checked their similarity to the full alignment and the inverted residue trees (Fig. 4C). We observed a decrease in tree similarity to the full tree proportional to the decrease in alignment length. However, the inverted residue tree distance is statistically less similar to the full alignment tree than the bootstraps, even among the bootstraps with an equivalent length of 20 residues (student t-test p-value: 0.0019). This result suggests that the inverted residues are just a minority of the sites in the alignment and that they disperse the phylogenetic information faster than the average. This view is compatible with the hypothesis that a few functional mutations are hidden behind an overwhelming amount of neutral or nearly-neutral variants, complicating their detection.

**Comparison of DIRphy inverted residues to Specificity Determining Sites (SDS) prediction**

We predicted the SDS of fish EGFR using SPEER (Chakraborty, et al. 2012), and observed a marked difference in the type of identified positions compared to DIRphy. To compare the two methods, we defined a “matching probability” as the dot product of the amino acid emission

probability vectors of two HMM models. Then, we compared, for each site of the fish EGFR alignment, the mean matching probability between HMM models of the orthologs (cypriniformes EGFRa to other fish EGFRa, and cypriniformes EGFRb to other fish EGFRb) with the mean matching probability between the paralogs (cypriniformes EGFRa to other fish EGFRb, and cypriniformes EGFRb to other fish EGFRa) (Fig. 4D). From this analysis, it is evident that SPEER tends to predict sites with a high ortholog-low paralog correlation. These sites are likely to possess a paralog-specific function. On the other hand, DIRphy identifies diametrically opposite sites, with a high paralog-low ortholog correlation. Both types of sites deviates from the diagonal, a pattern that suggests a functional adaptation. However, the sites that show an inversion are more challenging to identify for SDS predictions because, for a subset of species, the conservation pattern is inverted, and the signal is averaged out. In summary, the identification of residue inversion event has the potential to improve functional residue predictions, as DIRphy is able to identify functional sites that are overlooked by SDS prediction methods.

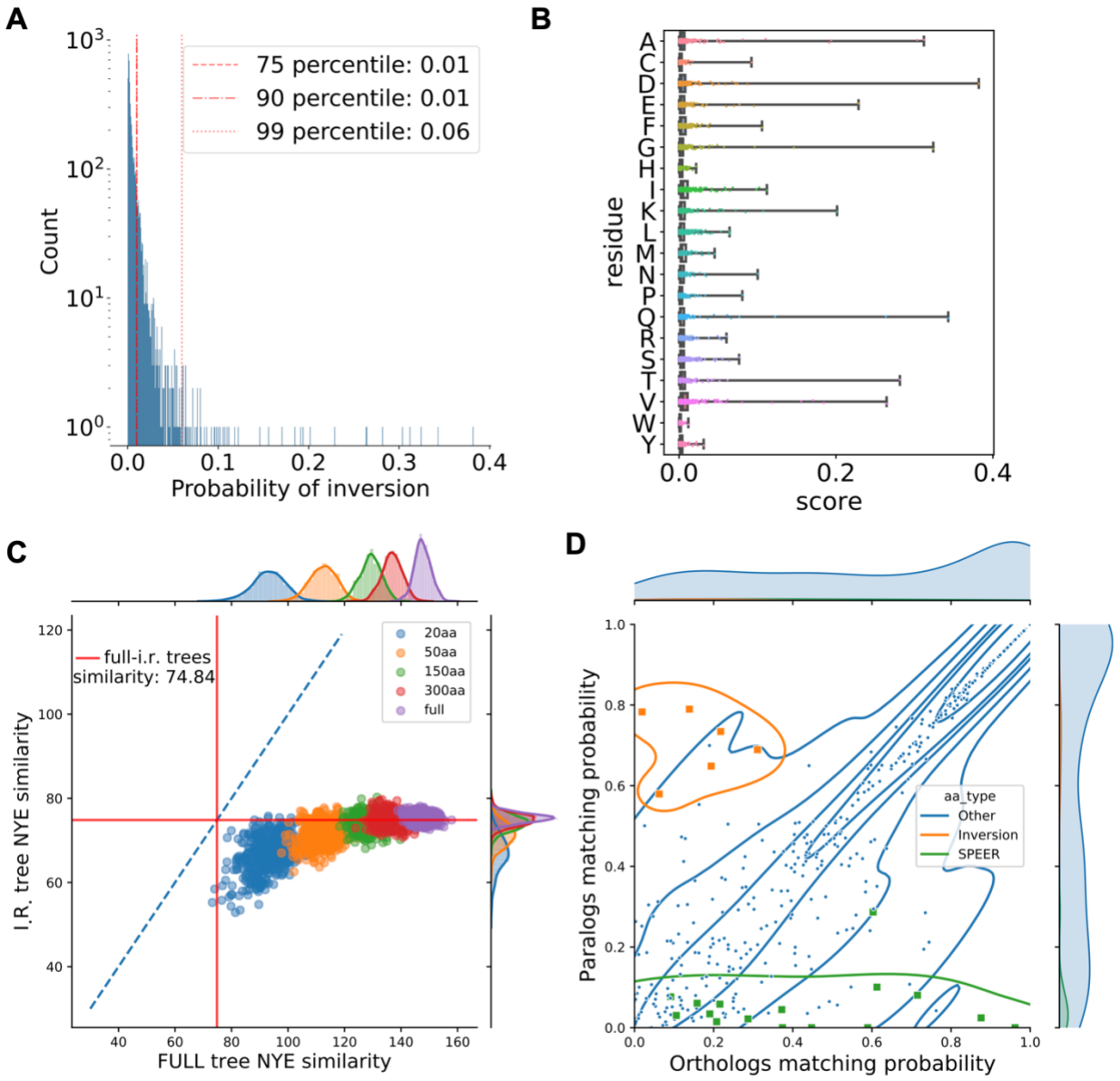



Fig. 4. Outcome to the validations. (A) Distribution of the simulated evolution DIRphy scores. A 5000 random amino acid sequence was evolved through the fish EGFR phylogenetic tree using the same evolutionary models used to generate the tree. The resulting MSA was used to compute the DIRphy score. (B) Distribution per amino acid of the simulated evolution DIRphy scores as shown by the reference (*S anshuiensis* EGFRa). (C) Bootstrap trees similarity to the full and inverted residue trees. The color represents the length of the sub-alignment used to generate the

bootstrap tree. The red line shows the similarity between the full and inverted residue trees. The blue line is the identity line. (D) Comparison of the sites identified by DIRphy and SPEER. The matching probability of the HMM of four sub-alignments was used to compare between species the orthologs (EGFRa vs EGFRa, EGFRb vs EGFRb) and paralogs (EGFRa vs EGFRb). The matching probability is calculated as the average of two dot products of the frequency arrays. The orange color shows sites where a residue inversion was identified, while the green color shows sites where the p-value of SPEER score is lower than 0.01.

In conclusion, we have detected an event in paralogs that lead to the inversion of functional residues. This new event has been described by a theoretical model and validated by literature and bioinformatics studies. To the best of our knowledge, we are the first to identify and describe the residue inversion event, possibly because this event might be rare. However, residue inversions are potentially exploited for functional divergence and, if missed, might lead to the wrong classification of proteins in the correct functional groups.

**Acknowledgments**

We thank Federica di Palma, Tarang K. Mehta, and Wilfried Haerty for the thoughtful discussions on fish phylogeny. Stanisław Dunin-Horkawicz and Dan Kozome for critical reading of the manuscript. We are grateful for the help and support provided by the Scientific Computing and Data Analysis section of Research Support Division at OIST. Financial support by the Okinawa Institute of Science and Technology to P.L. is gratefully acknowledged.

**Author contributions**

S.P. and P.L. designed the project. S.P. performed the research. S.P. and P.L. analyzed the data and wrote the manuscript.

## Materials and Methods

### Theoretical model

The model was built on the following assumptions: 1) equal branch length between the two paralogs: b1 = b2, b3 = b4 = b5 = b6; 2) only zero to one mutation can occur in each of the six branches; 3) after a mutation, each residue is equiprobable; 4) no selective pressure; 5) the probability of a mutation on a branch solely depends on the branch length (mutation rate) and is $P = 1 - e^{-\lambda}$ where P is the probability of a mutation and $\lambda$ is the mutation rate.

Given the probability of a mutation in each of the six branches, we can calculate the probability of all the 64 ($2^6$) configurations of mutations on the tree. A configuration unequivocally leads to a determined leaf node state (Supplementary Figure 4). We defined the leaf node states with the allowed values zero, one, or two. The state value represents the number of mutations that happened in the branches connecting to the leaf node (Fig. 1A). Then, we defined the probability of a match between leaf nodes based on the state value. The probabilities describe the situation in which a residue can mutate to one of 19 possible other residues. In some cases, the matching probability depends on the state of the ancestral node before speciation, e.g., a single or double mutation in the inner branch leaves. Finally, we defined seven categories based on the type of matching at the leaf nodes, and we defined their probability of occurrence based on the matching probabilities between leaf nodes.

**Model categories**

Here we give a brief description of the model categories and formulas used to calculate their probabilities in the model.

*Conserved*

$$P_{cons} = mi_{Xa,Ya} * mi_{Xb,Yb} * \max(mo_{Xa,Xb}, mo_{Xa,Yb}, mo_{Ya,Xb}, mo_{Ya,Yb})$$

Where *mi* stands for the inner match, *mo* stands for the outer match and corresponds to the probabilities of a match in figure 1B. The conserved category collects the states where a site is invariant in all four leaf nodes. It could arise from no mutations, but also from (two, three, or) four mutations to the same amino acid. In the formula, the maximum value of the outer matches is used as the best approximation of the conditional probability of matching all leaf nodes given the two inner matches.

*Type 1 Divergence*

$$P_{type1} =$$

$$= mi_{Xa,Ya} * (1 - mi_{Xb,Yb}) * (1 - \min(mo_{Xa,Xb}, mo_{Ya,Xb})) * (1 - \min(mo_{Xa,Yb}, mo_{Ya,Yb}))$$

$$+(1 - mi_{Xa,Ya}) * mi_{Xb,Yb} * (1 - \min(mo_{Xa,Xb}, mo_{Xa,Yb})) * (1 - \min(mo_{Ya,Xb}, mo_{Ya,Yb}))$$

The type 1 divergence collects the states where the amino acid is matching between species only in one paralog, while there is no match in the other paralog.

*Type 2 Divergence*

$$P_{type2} = mi_{Xa,Ya} * mi_{Xb,Yb} * (1 - P_{cons})$$

The type 2 divergence collects the states where the two paralogs display a different amino acid but are conserved between species. Type 1 and type 2 classifications are based on (Gu 2001).

*Recent Divergence*

$$P_{recent} = sum(mi_{Xa,Ya} * mo_{Xa,Xb} * (1 - mo_{Xb,Yb})$$

$$mi_{Xa,Ya} * mo_{Xa,Yb} * (1 - mo_{Xb,Yb})$$

$$mi_{Xb,Yb} * mo_{Xa,Xb} * (1 - mo_{Xa,Ya})$$

$$mi_{Xb,Yb} * mo_{Ya,Xb} * (1 - mo_{Xa,Ya}))$$

The recent divergence collects the states where only one leaf node is different (diverged) compared to the other three nodes.

*Inversion / Specie-Specific Adaptation*

$$P_{inv} = mo_{Xa,Yb} * mo_{Xb,Ya} * (1 - \min(mi_{Xa,Ya}, mi_{Xb,Yb}))$$

And similarly,

$$P_{ssa} = mo_{Xa,Xb} * mo_{Ya,Yb} * (1 - \min(mi_{Xa,Ya}, mi_{Xb,Yb}))$$

These two categories represent the states where the amino acid does not match between orthologs but matches between paralogs for the inversion or between species for the species-specific adaptation.

**Calculation of inverted residue score**

We devised a score to identify inverted residues in a phylogeny. The score was based on the probability of observing an inversion in the previously described model and calculated with the following steps: 1) Divide an MSA into four sub-alignments (two EGFR copies and two species groups). 2) Generate four amino acid frequency arrays, optionally normalized by pseudo-counts*. 3) Calculate the probability of a match between two groups using the dot product of the frequency array. 4) Calculate the joint probability of inversion (or similarly for species-specific adaptation) from the conditional probabilities and the frequency array matching using the following formulas:

$$P_{\overline{(M(Xa,Xb,Ya,Yb)}|M(Xa,Yb),M(Ya,Xb)} = \sum_{i}^{20} \frac{Xa[i] * Yb[i]}{M(Xa,Yb)} * (1 - \frac{Ya[i] * Xb[i]}{M(Ya,Xb)})$$

$$P_{\overline{(M(Xa,Xb,Ya,Yb)},M(Xa,Yb),M(Ya,Xb)} = P_{\overline{(M(Xa,Xb,Ya,Yb)}|M(Xa,Yb),M(Ya,Xb)} * M(Xa,Yb) * M(Ya,Xb)$$

Where *Xa*,*Xb*,*Ya*,*Yb* are the amino acid frequency arrays for the MSA sequence groups with the same names, *i* is the counter spanning each amino acid. The probability of a match (e.g., *M(Xa,Yb)*) is given by the array dot product: $M(Xa,Yb) = \sum_{i}^{20} Xa[i] * Yb[i]$. The latter joined probability represents the inverted residue score.

*** Pseudo-count normalization**

To correct any possible bias given by groups with a small number of species, we implemented a pseudo-count normalization of the amino acid frequency array, as previously done in Tatsuov *et al.* (Tatusov, et al. 1994). We used the LG protein substitution matrix (Le and Gascuel 2008) as background amino acid frequency probability. The value of the beta parameter for the pseudo-counts formula was set by default to five; however, it is possible to modify this parameter before running the pipeline.

**Fish dataset**

To test the inverted residue score, we generated a fish genome dataset. First, we downloaded all genomes from the European Nucleotide Archive (ENA) belonging to taxon 41665 (Actinopterygii). Through this method, we obtained 88 fish genomes. Next, we downloaded the pre-annotated fish EGFR protein sequences from the ENSEMBL database (Zerbino, et al. 2018) and used them to build an HMM profile with the HMMER package (Eddy 2011). The HMM profile was used as a query in Augustus package suite (Stanke, et al. 2006) to search for EGFR

related genes in the fish genomes. We then filtered out interrupted CDS, sequences clustering with other ErbBs in a phylogenetic tree, and sequences with an aberrant branch length in the non-synonymous codon tree. After this procedure, we obtained 167 fish EGFR protein sequences.

**Phylogenetic analysis**

We performed the phylogenetic analysis of the fish EGFR protein sequences using MAFFT (Katoh and Standley 2013) to align the sequences, and IQTREE (Nguyen, et al. 2015) with ModelFinder (Kalyaanamoorthy, et al. 2017) to search for the best evolutionary model to generate the phylogenetic tree. To generate the synonymous tree, we used paml CODEML (Yang 2007).

**Sequence and structure analysis**

The sequences and alignments were handled using Unipro Ugene (Okonechnikov, et al. 2012). The protein structure images and analyses were performed with UCSF Chimera (Goddard, et al. 2018). The modelling of fish EGFR structures was performed using the SWISS-MODEL web server (Waterhouse, et al. 2018) and AlphaFold 2 (Jumper, et al. 2021).

**Simulated evolution**

We ran a simulated evolution experiment using an in-house pipeline based on the Pyvolve python package (Spielman and Wilke 2015). The pipeline simulated an evolutionary pathway of 5000 random amino acids on the fish EGFR phylogenetic tree, using the same model of evolution and evolutionary rates that were used to construct the tree (JTT with rate

heterogeneity) (Jones, et al. 1992; Yang 1995; Soubrier, et al. 2012). The simulation generated an output alignment that was used to run the DIRphy pipeline, to calculate the base probability of an inversion.

**Bootstrap**

To perform the bootstrap, we used an in-house Matlab script. We selected from the fish EGFR DIRphy prediction the 19 sites with a score higher than the 99$^{th}$ percentile of the simulated evolution scores. We calculated a phylogenetic tree for the full alignment and the inverted residues alignment using the neighbor-joining algorithm (Saitou and Nei 1987) and the BLOSUM80 matrix (Henikoff and Henikoff 1992). The similarity distance between trees was calculated using the method described in Nye *et al.* (Nye, et al. 2006). We then performed 500 bootstrap alignments for each set of alignment lengths: full, 250, 100, 20. We excluded columns with 90% or more gaps and repeated the sampling whenever one sequence did not have at least one non-gap position. For each bootstrap alignment, we generated a tree, then calculated the distance to the full tree and inverted residue tree.

**References**

Abhiman S, Sonnhammer EL. 2005. Large-scale prediction of function shift in protein families with a focus on enzymatic function. Proteins-Structure Function and Bioinformatics 60:758-768.

Amitai G, Shemesh A, Sitbon E, Shklar M, Netanely D, Venger I, Pietrokovski S. 2004. Network analysis of protein structures identifies functional residues. J Mol Biol 344:1135-1146.

Amores A, Suzuki T, Yan YL, Pomeroy J, Singer A, Amemiya C, Postlethwait JH. 2004. Developmental roles of pufferfish Hox clusters and genome evolution in ray-fin fish. Genome Res 14:1-10.

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:W344-W350.

Bergthorsson U, Andersson DI, Roth JR. 2007. Ohno's dilemma: evolution of new genes under continuous selection. Proc Natl Acad Sci U S A 104:17004-17009.

Capra JA, Singh M. 2008. Characterization and prediction of residues determining protein functional specificity. Bioinformatics 24:1473-1480.

Casari G, Sander C, Valencia A. 1995. A method to predict functional residues in proteins. Nat Struct Biol 2:171-178.

Chakraborty A, Chakrabarti S. 2014. A survey on prediction of specificity-determining sites in proteins. Briefings in Bioinformatics 16:71-88.

Chakraborty A, Mandloi S, Lanczycki CJ, Panchenko AR, Chakrabarti S. 2012. SPEER-SERVER: a web server for prediction of protein specificity determining sites. Nucleic Acids Res 40:W242-248.

Dolinski K, Botstein D. 2007. Orthology and functional conservation in eukaryotes. Annu Rev Genet 41:465-507.

Eddy SR. 2011. Accelerated Profile HMM Searches. PLoS Comput Biol 7:e1002195.

Fischer JD, Mayer CE, Söding J. 2008. Prediction of protein functional residues from sequence by probability density estimation. Bioinformatics 24:613-620.

Goddard TD, Huang CC, Meng EC, Pettersen EF, Couch GS, Morris JH, Ferrin TE. 2018. UCSF ChimeraX: Meeting modern challenges in visualization and analysis. Protein Science 27:14-25.

Gomez A, Volff JN, Hornung U, Schartl M, Wellbrock C. 2004. Identification of a second egfr gene in Xiphophorus uncovers an expansion of the epidermal growth factor receptor family in fish. Mol Biol Evol 21:266-275.

Gout JF, Lynch M. 2015. Maintenance and Loss of Duplicated Genes by Dosage Subfunctionalization. Mol Biol Evol 32:2141-2148.

Gu X. 2001. Maximum-likelihood approach for gene family evolution under functional divergence. Mol Biol Evol 18:453-464.

Gu X, Zou Y, Su Z, Huang W, Zhou Z, Arendsee Z, Zeng Y. 2013. An update of DIVERGE software for functional divergence analysis of protein family. Mol Biol Evol 30:1713-1719.

Hannenhalli SS, Russell RB. 2000. Analysis and prediction of functional sub-types from protein sequence alignments. J Mol Biol 303:61-76.

Henikoff S, Henikoff JG. 1992. Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci U S A 89:10915-10919.

Hubbard SR, Miller WT. 2007. Receptor tyrosine kinases: mechanisms of activation and signaling. Curr Opin Cell Biol 19:117-123.

Huerta-Cepas J, Dopazo J, Huynen MA, Gabaldón T. 2011. Evidence for short-time divergence and long-time conservation of tissue-specific expression after gene duplication. Brief Bioinform 12:442-448.

Innan H, Kondrashov F. 2010. The evolution of gene duplications: classifying and distinguishing between models. Nature Reviews Genetics 11:97-108.

Jones DT, Taylor WR, Thornton JM. 1992. The rapid generation of mutation data matrices from protein sequences. Comput Appl Biosci 8:275-282.

Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, et al. 2021. Highly accurate protein structure prediction with AlphaFold. Nature 596:583-589.

Kalinina OV, Mironov AA, Gelfand MS, Rakhmaninova AB. 2004. Automated selection of positions determining functional specificity of proteins by comparative analysis of orthologous groups in protein families. Protein Sci 13:443-456.

Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. Nature Methods 14:587-+.

Katoh K, Standley DM. 2013. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Molecular Biology and Evolution 30:772-780.

Kimura M. 1991. The neutral theory of molecular evolution: a review of recent evidence. Jpn J Genet 66:367-386.

Laisney J, Braasch I, Walter RB, Meierjohann S, Schartl M. 2010. Lineage-specific co-evolution of the Egf receptor/ligand signaling system. Bmc Evolutionary Biology 10:16.

Le SQ, Gascuel O. 2008. An improved general amino acid replacement matrix. Mol Biol Evol 25:1307-1320.

Lee D, Redfern O, Orengo C. 2007. Predicting protein function from sequence and structure. Nat Rev Mol Cell Biol 8:995-1005.

Leinonen R, Akhtar R, Birney E, Bower L, Cerdeno-Tárraga A, Cheng Y, Cleland I, Faruque N, Goodgame N, Gibson R, et al. 2011. The European Nucleotide Archive. Nucleic Acids Res 39:D28-31.

Li L, Stoeckert CJ, Jr., Roos DS. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res 13:2178-2189.

Lichtarge O, Bourne HR, Cohen FE. 1996. An evolutionary trace method defines binding surfaces common to protein families. Journal of Molecular Biology 257:342-358.

Lorin T, Brunet FG, Laudet V, Volff JN. 2018. Teleost Fish-Specific Preferential Retention of Pigmentation Gene-Containing Families After Whole Genome Duplications in Vertebrates. G3 (Bethesda) 8:1795-1806.

Macqueen DJ, Johnston IA. 2014. A well-constrained estimate for the timing of the salmonid whole genome duplication reveals major decoupling from species diversification. Proc Biol Sci 281:20132881.

Mallik S, Tawfik DS. 2020. Determining the interaction status and evolutionary fate of duplicated homomeric proteins. PLoS Comput Biol 16:e1008145.

McClintock JM, Kheirbek MA, Prince VE. 2002. Knockdown of duplicated zebrafish hoxb1 genes reveals distinct roles in hindbrain patterning and a novel mechanism of duplicate gene retention. Development 129:2339-2354.

Meyer A, Van de Peer Y. 2005. From 2R to 3R: evidence for a fish-specific genome duplication (FSGD). Bioessays 27:937-945.

Mirny LA, Gelfand MS. 2002. Using Orthologous and Paralogous Proteins to Identify Specificity-determining Residues in Bacterial Transcription Factors. Journal of Molecular Biology 321:7-20.

Mitchell MB. 1955. ABERRANT RECOMBINATION OF PYRIDOXINE MUTANTS OF Neurospora. Proc Natl Acad Sci U S A 41:215-220.

Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. Molecular Biology and Evolution 32:268-274.

Nye TM, Liò P, Gilks WR. 2006. A novel algorithm and web-based tool for comparing two alternative phylogenetic trees. Bioinformatics 22:117-119.

Oda K, Matsuoka Y, Funahashi A, Kitano H. 2005. A comprehensive pathway map of epidermal growth factor receptor signaling. Mol Syst Biol 1:2005.0010.

Ogiso H, Ishitani R, Nureki O, Fukai S, Yamanaka M, Kim JH, Saito K, Sakamoto A, Inoue M, Shirouzu M, et al. 2002. Crystal structure of the complex of human epidermal growth factor and receptor extracellular domains. Cell 110:775-787.

Okonechnikov K, Golosova O, Fursov M, team tU. 2012. Unipro UGENE: a unified bioinformatics toolkit. Bioinformatics 28:1166-1167.

Panchenko AR, Kondrashov F, Bryant S. 2004. Prediction of functional sites by analysis of sequence and structure conservation. Protein Sci 13:884-892.

Reva B, Antipin Y, Sander C. 2007. Determinants of protein function revealed by combinatorial entropy optimization. Genome Biol 8:R232.

Rockah-Shmuel L, Tóth-Petróczy Á, Tawfik DS. 2015. Systematic Mapping of Protein Mutational Space by Prolonged Drift Reveals the Deleterious Effects of Seemingly Neutral Mutations. PLoS Comput Biol 11:e1004421.

Saitou N, Nei M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4:406-425.

Sankararaman S, Kolaczkowski B, Sjölander K. 2009. INTREPID: a web server for prediction of functionally important residues by evolutionary analysis. Nucleic Acids Res 37:W390-395.

Soubrier J, Steel M, Lee MS, Der Sarkissian C, Guindon S, Ho SY, Cooper A. 2012. The influence of rate heterogeneity among sites on the time dependence of molecular rates. Mol Biol Evol 29:3345-3358.

Spielman SJ, Wilke CO. 2015. Pyvolve: A Flexible Python Module for Simulating Sequences along Phylogenies. PLoS One 10:e0139047.

Stanke M, Keller O, Gunduz I, Hayes A, Waack S, Morgenstern B. 2006. AUGUSTUS: ab initio prediction of alternative transcripts. Nucleic Acids Res 34:W435-439.

Storm CE, Sonnhammer EL. 2002. Automated ortholog inference from phylogenetic trees and calculation of orthology reliability. Bioinformatics 18:92-99.

Tadaki DK, Niyogi SK. 1993. The functional importance of hydrophobicity of the tyrosine at position 13 of human epidermal growth factor in receptor binding. J Biol Chem 268:10114-10119.

Tareen A, Kinney JB. 2020. Logomaker: beautiful sequence logos in Python. Bioinformatics 36:2272-2274.

Tatusov RL, Altschul SF, Koonin EV. 1994. Detection of conserved segments in proteins: iterative scanning of sequence databases with alignment blocks. Proc Natl Acad Sci U S A 91:12091-12095.

Tatusov RL, Koonin EV, Lipman DJ. 1997. A genomic perspective on protein families. Science 278:631-637.

Wang JT, Li JT, Zhang XF, Sun XW. 2012. Transcriptome analysis reveals the time of the fourth round of genome duplication in common carp (Cyprinus carpio). BMC Genomics 13:96.

Waterhouse A, Bertoni M, Bienert S, Studer G, Tauriello G, Gumienny R, Heer FT, de Beer TA P, Rempfer C, Bordoli L, et al. 2018. SWISS-MODEL: homology modelling of protein structures and complexes. Nucleic Acids Research 46:W296-W303.

Wicker N, Perrin GR, Thierry JC, Poch O. 2001. Secator: a program for inferring protein subfamilies from phylogenetic trees. Mol Biol Evol 18:1435-1441.

Wilkins AD, Bachman BJ, Erdin S, Lichtarge O. 2012. The use of evolutionary patterns in protein annotation. Curr Opin Struct Biol 22:316-325.

Wuster A, Venkatakrishnan AJ, Schertler GF, Babu MM. 2010. Spial: analysis of subtype-specific features in multiple sequence alignments of proteins. Bioinformatics 26:2906-2907.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol 24:1586-1591.

Yang Z. 1995. A space-time process model for the evolution of DNA sequences. Genetics 139:993-1005.

Yano S, Kondo K, Yamaguchi M, Richmond G, Hutchison M, Wakeling A, Averbuch S, Wadsworth P. 2003. Distribution and function of EGFR in human tissue and the effect of EGFR tyrosine kinase inhibition. Anticancer Res 23:3639-3650.

Zerbino DR, Achuthan P, Akanni W, Amode MR, Barrell D, Bhai J, Billis K, Cummins C, Gall A, Giron CG, et al. 2018. Ensembl 2018. Nucleic Acids Res 46:D754-d761.

Zmasek CM, Eddy SR. 2002. RIO: analyzing proteomes by automated phylogenomics using resampled inference of orthologs. BMC Bioinformatics 3:14.