****

**SCHOOL OF COMPUTER SCIENCE AND ENGINEERING**

**SCHOOL OF BIOTECHNOLOGY AND BIOMOLECULAR SCIENCES**

**Quantitative Effects of Residue Mutation on Protein Interaction Partners**

*Ricky Chang*

ricky.chang@student.unsw.edu.au

Thesis B Report submitted as a requirement for the degree

Bachelor of Engineering (Bioinformatics)

Supervisor: Prof. Marc Wilkins

Co-Assessor: Dr. Bruno Gaeta

Submitted: June 5th 2007

**CONTENTS**

[CHAPTER 1: INTRODUCTION 1](#_Toc168097509)

[1.1 Aims 2](#_Toc168097510)

[1.2 Differences from Thesis A Proposal 3](#_Toc168097511)

[CHAPTER 2: BACKGROUND 5](#_Toc168097512)

[2.1 PPI Network Topology 5](#_Toc168097513)

[2.2 Party/Date Hubs 7](#_Toc168097514)

[2.3 Definition of Ortholog vs. Paralog vs. Homolog 9](#_Toc168097515)

[2.3 Ortholog Hub Conservation Across Species 11](#_Toc168097516)

[2.4 Paralogous Gene Duplication 11](#_Toc168097517)

[2.5 Current Methods and Limitations 12](#_Toc168097518)

[2.5.1 Yeast 2 Hybrid Assay (Y2H) 12](#_Toc168097519)

[2.5.2 Mass spectrometry purification 14](#_Toc168097520)

[2.5.3 Synthetic lethality (knockout studies) 14](#_Toc168097521)

[2.5.4 Bioinformatic analysis 15](#_Toc168097522)

[CHAPTER 3: MATERIALS 16](#_Toc168097523)

[3.1 Datasets 16](#_Toc168097524)

[3.2.1 Databases 16](#_Toc168097525)

[3.2.2 Interaction Datasets 18](#_Toc168097526)

[3.2 Internet Retrieved Data 19](#_Toc168097527)

[3.2.1 iPfam Domain Residue Locations 19](#_Toc168097528)

[3.2.2 PDB Interacting Residue Locations 20](#_Toc168097529)

[3.3 Tools 21](#_Toc168097530)

[3.3.1 Multiple Alignment Programs 21](#_Toc168097531)

[3.2.3 R Statistical Package 22](#_Toc168097532)

[CHAPTER 4: METHODS & IMPLEMENTATION 23](#_Toc168097533)

[4.1 Programming Workflow 24](#_Toc168097534)

[4.2 Workflow Breakdown 25](#_Toc168097535)

[4.2.1 Intersection of Datasets 25](#_Toc168097536)

[4.2.2 Selection of KOGs 26](#_Toc168097537)

[4.2.3 Extraction of Domain Sequences 28](#_Toc168097538)

[4.2.4 Download of .sth 28](#_Toc168097539)

[4.2.5 Multiple Sequence Alignment of Targets 28](#_Toc168097540)

[4.2.6 Overlay of Interacting Residue Data 28](#_Toc168097541)

[4.2.7 Compilation of Scores and Statistics 28](#_Toc168097542)

[CHAPTER 5: EXPERIMENTS & RESULTS 30](#_Toc168097543)

[5.1 General Analysis 30](#_Toc168097544)

[5.2 Investigation of paralog domain similarity 33](#_Toc168097545)

[5.2 Investigation of new PPI partners 36](#_Toc168097546)

[5.3 Investigation of new protein domains 37](#_Toc168097547)

[5.4 Investigation of paralogs 37](#_Toc168097548)

[5.6 Investigation of mutation on partner loss 39](#_Toc168097549)

[5.5 Investigation of Robustness 43](#_Toc168097550)

[CHAPTER 6: CONCLUSION 44](#_Toc168097551)

[6.1 Interpretation of Results 44](#_Toc168097552)

[6.2 Limitations 44](#_Toc168097553)

[6.3 Future Work 45](#_Toc168097554)

[BIBLIOGRAPHY 49](#_Toc168097555)

## 1.2 Differences from Thesis A Proposal

The investigations outlined in the Thesis A proposal have been changed significantly, in consideration of the lack of supporting data for yeast-to-human paralog/ortholog proteins. Consequently, sections 3 (MATERIALS) and 5 (EXPERIMENTS & RESULTS), as well as the thesis title, have been modified to reflect this.

However, new insight has been gained since the completion of the project implementation. Yeast-to-human protein evolution is address in future work, under section 6.3.

## 2.3 Definition of Ortholog vs. Paralog vs. Homolog

Homolog

* A gene related to a second gene by descent from a common ancestral DNA sequence. The term, homolog, may apply to the relationship between genes separated by the event of speciation (see ortholog) or to the relationship betwen genes separated by the event of genetic duplication (see paralog).

Ortholog

* Orthologs are genes in different species that evolved from a common ancestral gene by speciation. Normally, orthologs retain the same function in the course of evolution. Identification of orthologs is critical for reliable prediction of gene function in newly sequenced genomes. (See also Paralogs.).

Paralog

* Paralogs are genes related by duplication within a genome. Orthologs retain the same function in the course of evolution, whereas paralogs evolve new functions, even if these are related to the original one.

[Gogarten] retrieve paper and expand

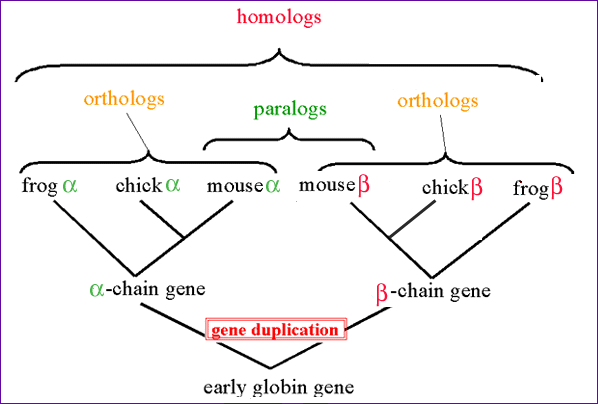


Fig. An illustration of the relationship between homologs, orthlogs and paralogs by example of the globin gene family [5].

**Homologous sequences.** Orthologs and Paralogs are two types of homologous sequences. Orthology describes genes in different species that derive from a common ancestor. Orthologous genes may or may not have the same function. Paralogy describes homologous genes within a single species that diverged by gene duplication.

## 3.2.2 PDB Interacting Residue Locations

***DISCOVERY OF DOMAIN-DOMAIN INTERACTIONS***

Upon further literature search, it has been discovered that iPfam associations are from domain-domain interactions between PDB entries, which were experimentally determined [9]. Our source of domain-domain interaction information consisted of several datasets [10-12]. As expected, those reported from iPfam were found, and the interacting residues were found for those interactions. For the Lee and Guimarães datasets, this information was not available due to their absence in iPfam. If their predictions are correct, these domain-domain interactions would add to the ones already available in iPfam. It would be worthwhile to perform a literature search on experiments to validate these predicted interactions.

***NATURE OF INTERACTING RESIDUES BETWEEN DOMAINS***

Our principle focus is on interactions between PfamA domains. The method of determining interacting residues in iPfam is to map Pfam domains onto PDB structures. The exact amino acid residues involved in the DDI can be determined with interdomain bonds. Currently, there is no differentiation between biological and crystal contacts of these bonds [10]. In this project, the interacting domains of orthologs from our own dataset have been aligned against the domains provided, when this data is available. It is our assumption that the interacting residue locations have been preserved.

## 3.3 Tools

### 3.3.1 Multiple Alignment Programs

3 different multiple alignment programs were trialed – these were Clustalw [13], T-Coffee [14] & ProbCons [15].

Our targets of alignment consist of domains from several KOG [16] confirmed yeast orthologs, and another few domains from orthologs retrieved from iPfam which have experimentally-derived specific DDI’s mapped onto individual residues. These iPfam orthologs tend to come from another species (most commonly Homo Sapiens). A major concern of performing an alignment across such distantly related eurkaryotes is the accuracy of the alignment due to widely divergent mutations. Of the 3 programs used, ProbCons provided the best alignment, upon visual inspection of the output in Jalview [17].

An indicator of performance can be given in the case study of PF00069~PF00023. All 3 programs did fairly well in a region of >40% “high-confidence” similarity. The difference is marked when there are alternating “high-confidence” and “low-confidence” regions. At the boundary of these regions, all programs did make some wrong assignments of gaps and alignments, before returning back to the next “high-confidence” region. However, ProbCons tended to make better decisions inside the uncertain “low-confidence” regions. Thus, ProbCons was used in the primary domain alignment, despite its lack of features compared to more extensive packages such as T-Coffee. It is believed that its alignments will introduce fewer systematic errors, and enhance the quality of the end analysis. For further measures of similarity of the target sequences, features from T-Coffee/Clustalw have to be used instead.

An observation made at the end of this exercise is that there is a noticeable improvement from Clustalw, the most widely used multiple alignment software, to ProbCons. On the whole, alignment programs still need to be more robust in dealing with insertions (by assignment of proper gaps), which is prevalent as protein domains evolve from lower to higher organisms.

### 3.2.3 R Statistical Package

R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS.

Use of R in this project include Scatter-plots, Q-Q plots, conditioning-plots)

More information about R can be found at http://www.r-project.org/. [29]

# CHAPTER 4: METHODS & IMPLEMENTATION

In this section, methods and implementation of the project will be described in detail. Any assumptions and/or rationalizations are accompanied by a justification of why it was made.

## 4.1 Programming Workflow

INTERNET CONNECTION

KOG

Domain Sequences

PPI

merge\_ppi

remove

\_singlish

make\_a~b

insert\_ddi\_stats

generate\_

fasta

perform\_

msa

retrieve\_sth

extract\_

domain

PDB IR files

parse\_ir

make\_

scores

### 4.2.2 Selection of KOGs

The goal of this step is to eliminate as much irrelevant and un-interpretable data as possible before proceeding on to subsequent internet connection-required methods.

For a KOG to be eligible for consideration, it requires ≥1 common protein partners between any 2 KOG members. This is best illustrated by an example:

PFA~PFB-KOG1234

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| SwissProt  Accession | **sB1** | **sB2** | **sB3** | **sB4** | **sB5** | **sB6** | **n** |
| **sA1** |  |  | X | x |  | x | 3 |
| **sA2** | X |  |  | X |  |  | 2 |
| **sA3** | x |  | X |  |  |  | 2 |
| **sA4** |  | X |  |  | X | x | 3 |

X – indicates PPI between sA <=> sB

COMMON = 4

TOTAL = 6

COMMON counts unique common partners between any 2 sA’s. A more intuitive counting of COMMON would be to count common partners between all KOG members. The latter method is not applied due to the low density of data. COMMON is used as a threshold for defining a subset within AC, known as CC. CC members tend to have better alignments, more interacting partners, less error-prone, and thus having higher confidence intervals.

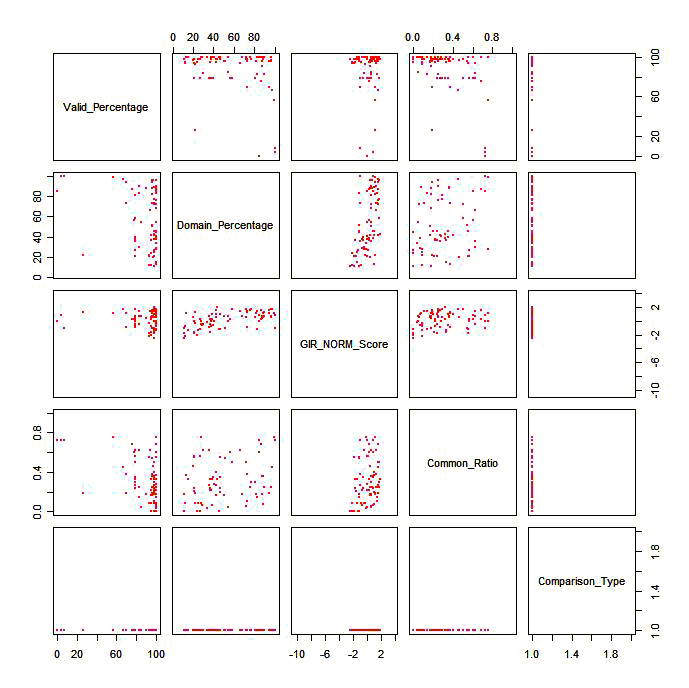
Additionally, any KOG with COMMON < 1 are discarded, i.e., KOGs with no common partners between any members, or single member KOGs.

TOTAL simply counts the total number of unique partners that KOG members have, regardless of commonality.

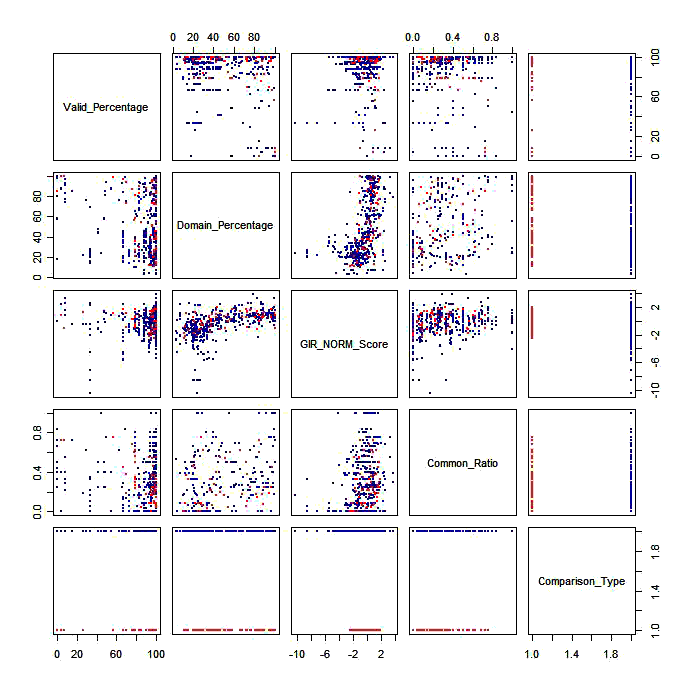
# CHAPTER 5: EXPERIMENTS & RESULTS

Due to the short time span of this project, it is important to formulate a series of focused and testable hypotheses which yield measurable results. From these specific questions, more conclusions can be drawn about the nature of domain mutations on interacting residues, and its effect on partner loss.

## 5.1 General Analysis



CC Dataset



AC Dataset

## 5.2 Investigation of paralog domain similarity

Hypothesis

As the percentage of mutations between paralog domains increases, the similarity between their interacting residues also increases.

Method

Results

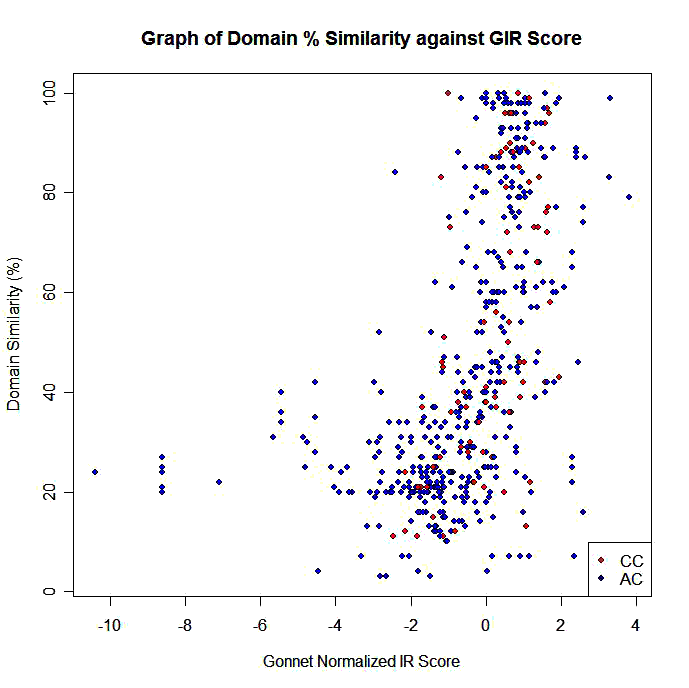


Fig. An (x,y) plot shows a positive correlation between domain similarity and GIR score.

An (x,y) plot is sufficient for a positive relationship between domain similarity and GIR score to be observed. However, this is an expected result, if no distinction is made between domain residues and interacting residues.

Thus, the original hypothesis can be expanded:

As domains mutate and diverge in sequence, their interacting residues mutate at a rate independent of that of the domain.

## 5.6 Investigation of mutation on partner loss

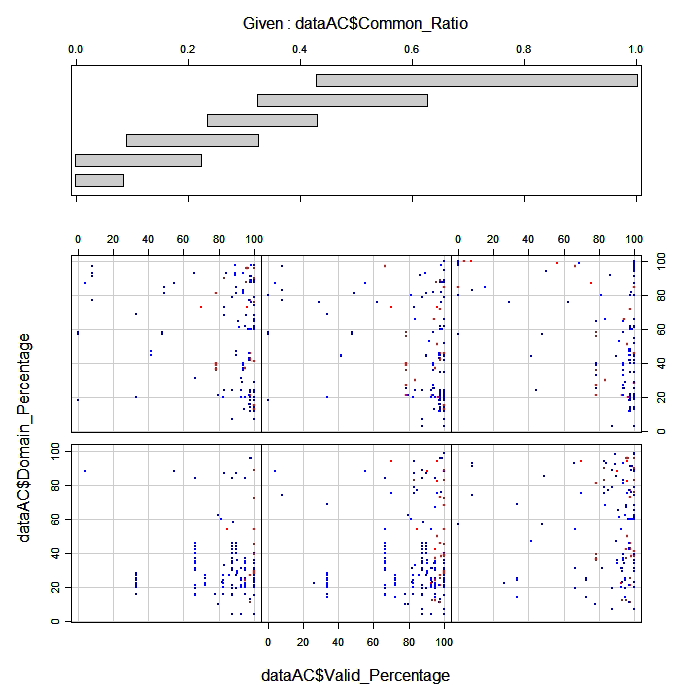


Fig. Conditional Plot of Domain Similarity % against Valid Residues %, partitioned based on common ratio standard deviations.

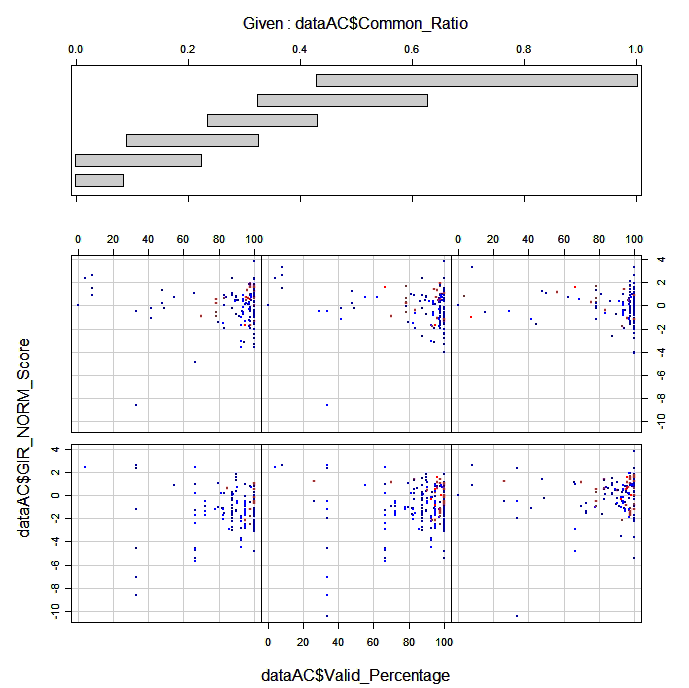


Fig. Conditional Plot of GIR Score against Valid Residues %, partitioned based on common ratio standard deviations. Clustering of high valid % due to good multiple alignment.

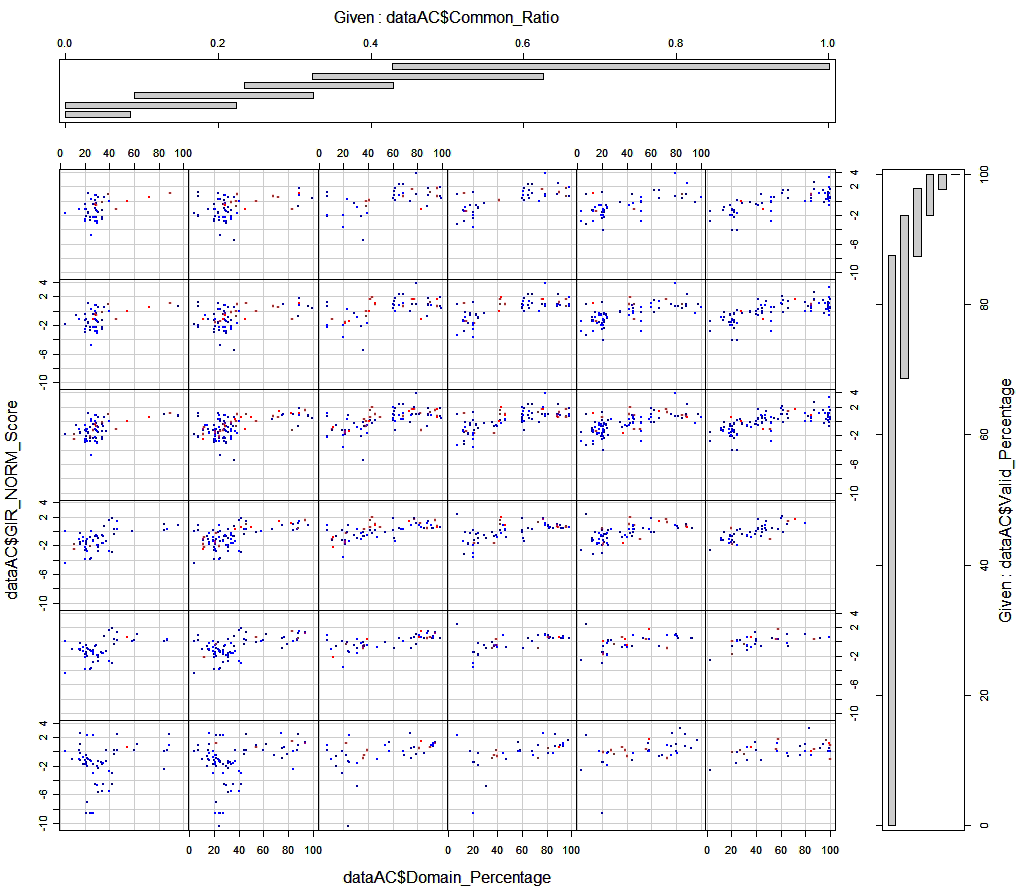


Fig. Conditional Plot of GIR Score against Valid Residues %, partitioned based on common ratio and valid residue % standard deviations. Clustering of high valid % due to good multiple alignment.

There is a noticeable cluster of values of low common ratio, moderate to high valid %, neutral GIR Scores and low to moderate domain similarity %. Positive correlation for partitions of moderate to high common ratios and valid residue %

## 6.2 Limitations

The weakest link of the programming workflow is the inaccuracy of the domain multiple alignments. Even with ProbCons, the best MSA program at the moment, many misalignments are made with simple inserts and deletions, thus adding frameshift to the actual homologous regions.

Sparseness of data, reliability

Biased PPI

Under-representation of membrane prot

## 6.3 Future Work

Inparanoid

* Best-best hits between sequences from 2 different species
* 2 MAIN ORTHOLOGS – defining members of an orthologous group
* Others added if they are closely related to main orthologs, “IN-PARALOGS”
* Division of sequences into orthologous groups is a measure which can be varied.

Nature of gene duplication – in or out paralogs – leading to Cell localization