



Jaypee University of Information Technology
Solan (H.P.)
LEARNING RESOURCE CENTER

Acc. Num. SP04100 Call Num:

General Guidelines:

- ◆ Library books should be used with great care.
- ◆ Tearing, folding, cutting of library books or making any marks on them is not permitted and shall lead to disciplinary action.
- ◆ Any defect noticed at the time of borrowing books must be brought to the library staff immediately. Otherwise the borrower may be required to replace the book by a new copy.
- ◆ The loss of LRC book(s) must be immediately brought to the notice of the Librarian in writing.

Learning Resource Centre-JUIT



SP04100

**DESIGNING A TOOL TO DETERMINE THE
FUNCTIONAL RESIDUES OF ANY PROTEIN
SEQUENCE USING THE PRINCIPLE OF
ORTHOLOGOUS AND PARALOGOUS GROUPS**

By

**ROHIT RAJEEV MARWAH – 041541
SOM SHANKAR SINGH – 041542**



MAY-2008

**Submitted in partial fulfillment of the Degree of Bachelor of
Technology**

**DEPARTMENT OF BIOINFORMATICS AND
BIOTECHNOLOGY
JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY
WAKNAGHAT**

CERTIFICATE

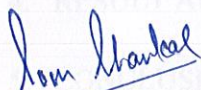
This is to certify that the work entitled, "DESIGNING A TOOL TO DETERMINE THE FUNCTIONAL RESIDUES OF ANY PROTEIN SEQUENCE USING THE PRINCIPLE OF ORTHOLOGOUS AND PARALOGOUS GROUPS" submitted by ROHIT RAJEEV MARWAH (041541) and SOM SHANKAR SINGH (041542) in partial fulfillment for the award of degree of Bachelor of Technology in Bioinformatics of Jaypee University of Information Technology has been carried out under my supervision. This work has not been submitted partially or wholly to any other University or Institute for the award of this or any other degree or diploma.

Chait
15/05/08

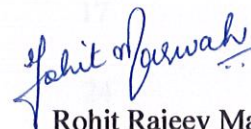
DR. CHITTARANJAN ROUT
Senior Lecturer
Department of Biotechnology and Bioinformatics
Jaypee University of Information Technology
Waknaghat, Solan-173215
Himachal Pradesh

ACKNOWLEDGEMENT

We would like to extend our heartfelt thanks and reverence to our project teacher **Dr. C. ROUT** for his continuous support in our project. He has been an inspiration for us in being innovative and liberated. He showed us different ways to approach a research problem and the need to be persistent for accomplishing any goal. He has always been there to discuss our ideas to proofread and asked thought provoking questions helping us think through our problems.



Som Shankar Singh



Rohit Rajeev Marwah

TABLE OF CONTENTS

<u>TOPICS</u>	<u>PAGE NO.</u>
1. ABSTRACT	8
2. INTRODUCTION	9
3. MATERIALS AND METHODS	12
4. RESULT AND DISCUSSION	17
5. CONCLUSION	24
6. FUTURE PROSPECTIVE	25
7. BIBLIOGRAPHY	26

LIST OF ABBREVIATIONS

DNA – Deoxyribonucleic Acid

GH – Growth Hormone

hGH – Human Growth Hormone

MI – Mutual Information

MSA – Multiple Sequence Alignment

NCBI– National Center for Biotechnology Information

P(I) – Statistical significance of MI

PL – Placental Lactogen

PRL – Prolactin

SGH – Somatotropin Growth Hormone

LIST OF FIGURES AND TABLES

FIGURES

1. Figure 1: Steps followed in this project. Page no. 11
2. Figure 2: Output of MSA of the individual orthologous groups i.e. (A) SGH, (B) PL and (C) PRL done through ClustalW. Page no. 17
3. Figure 3: Profile alignment between three paralogous groups SGH, Pro, and PL. Page no. 18
4. Figure 4: Shows the observed I_i and $P(I)$ for three orthologous groups taken together calculated by three methods: (A) Shuffling, (B) Bootstrapping and (C) Randomization. Page no. 20
5. Figure 5: Shows the screen shot of the tool. Page no. 21

TABLES

1. Table 1: Showing Functional Residues in three orthologous groups (SGH, PRL and PL). Page no. 21,22

ABSTRACT

Functional residues serve a dominant role in mediating the cooperativity of the protein ensemble and determination of functional residues of any protein has become an important factor. The basis of functional specificity of proteins is assumed to be conserved among orthologs and is different among paralogs. We used this assumption to identify residues which determine specificity of any protein family. Finding such residues is crucial for understanding mechanisms of molecular recognition and for rational protein and drug design.

Assuming conservation of specificity among orthologs and different specificity of paralogs, we identify residues that correlate with this grouping by specificity. So we design a tool which determines the functional residues of any protein sequence using the principle of orthologs and paralogs. In this project we used somatotropin growth hormone as a test sequence and then we found out its paralogs namely prolactin and placental lactogen which have their corresponding orthologs. Multiple sequence alignment (MSA) was done for individual groups: and profile alignment of the above mentioned individual groups is performed. Mutual information (MI) is calculated column wise. The main part of this tool is to compute statistical significance¹ of the MI values using three methods like, shuffling¹; random orthologs group generation¹; and replicate dataset by parametric bootstrap². Residues in columns of the MSA which gives high MI and low P(I) values are considered functional residue. Available programs, some created programs and some in house programs are combined to design the complete tool.

INTRODUCTION

The project is about designing a tool that would help the user to find the functional residues of any protein using orthologs and paralogs. This project undertakes various approaches together to make one final tool. This tool will be one final destination for its user as now he will not have to use different tools for different tasks involved in determining the functional residues.

The concepts of orthology and paralogy were originally introduced by Walter Fitch in 1970^{3,4} and recently became a subject of active discussion. Briefly, orthologs are genes in different organisms which are direct evolutionary counterparts of each other. Both paralogs and orthologs are assumed to have similar general biochemical functions, while orthologs are also believed to have the same specificity. In this study we exploit another property of orthologs: similar specificity, as contrasted by different specificities of paralogs.

If the above assumption is correct, grouping by orthology becomes grouping of proteins by specificity. Here we developed a method, which uses such grouping to identify amino acid residues that determine the protein specificity. Specificity determining residues can be very hard to find even when the structure of a protein or a complex is available, since very few amino acid residues provide specific recognition. Computational prediction of the specificity determinants can substantially reduce experimental efforts and provide guidance for rational re-design of protein function.^{5,6}

The idea of our method is (1) to start from a family of paralogs in one genome, find orthologs for each member of the family in other genomes, (2) forming different groups containing paralog and its orthologs respectively, and (3) apply statistical method to identify residues that are functionally important and can better discriminate between orthologous (specificity) groups.

In our case firstly, we have taken three orthologous groups i.e. somatotropin growth hormone, prolactin and placental lactogen because their both structure and function are known.

PRL, PL and GH are pituitary hormones that regulate an extensive variety of important physiological functions. While growth hormone biology generally centers around the regulation and differentiation of muscle, cartilage and bone cells, it is the PRL hormones and receptors that display a much broader spectrum of activities, ranging in diversity from their well-known effects in mammalian reproductive biology to osmoregulation in fishes and nesting behavior in birds⁷. An additional set of activities is induced by post-translationally modified forms of PRL and probably reacts through a noncytokine type of receptor⁸.

The biology of PRL and GH is integrated on many levels⁹; however, over the 400 million years since they diverged from a common gene parent, different regulating components have

evolved that distinguish them^{10,11}. In primates, the GH receptor (GHR) is activated solely by homodimerization through its cognate hormone^{10,12}. However, PRL biology works through regulated cross reactivity; most receptors are programmed to bind three hormones, PRL, PL and GH¹³.

SGH, commonly known as growth hormone (GH) is a polypeptide chain containing about 190 amino acid residues, produced by the pituitary gland in mammals and is responsible for a number of anabolic processes. Growth hormone is also of considerable interest as a drug. Secretion of SGH is regulated by two peptides that act to either stimulate (growth hormone-releasing factor (GRF)) or inhibit (somatostatin) release of SGH from the pituitary gland. Functions of growth hormone include: modify carcass composition, improve feed, efficiency enhance growth rate and milk yield.

PRL is a peptide hormone primarily associated with lactation. It is synthesised and secreted by lactotrope cells in the adenohypophysis (anterior pituitary gland). It is also produced in other tissues including the breast and the decidua. It stimulates the mammary glands to produce milk (lactation), provide the body with sexual gratification after sexual acts, stimulate proliferation of oligodendrocyte precursor cells, the surfactant synthesis of the foetal lungs at the end of the pregnancy and immune tolerance of the fetus by the maternal organism during pregnancy.

PL also called human chorionic somatomammotropin is a polypeptide placental hormone. Its structure and function is similar to that of human growth hormone. It modifies the metabolic state of the mother during pregnancy to facilitate the energy supply of the fetus. HPL is anti-insulin. HPL consists of 190 amino acids that are linked by two disulfite bonds and is secreted by the syncytiotrophoblast during pregnancy. PL affects the metabolic system of the maternal organism. HPL increases production of insulin and IGF-1 and increases insulin resistance and carbohydrate intolerance.

Secondly, we have taken 15 orthologs of SGH, 10 orthologs of PRL and 9 orthologs of PL. Three groups were formed containing SGH, PRL, PL and their orthologs respectively. Now profile alignment was done using ClustalW¹⁴, Firstly the profile alignment between the SGH and PRL was done. Secondly, the profile alignment was done between PL and the resulting MSA.

Thirdly, three different statistical methods were followed to identify residues that were functionally important and can better discriminate between orthologous (specificity) groups. These methods were shuffling, randomization and bootstrapping method respectively. These methods calculate the mutual information (MI) and then P(I) after which graph is plotted between MI and P(I). Our statistical procedure determines whether positions in the MSA can discriminate between functional sub-families better than the sequence similarity. Residues that satisfy these criteria are predicted to be specificity-determining. Thus the residues which are having higher MI values and lower P(I) values are the functionally important residues. (Steps followed in these project is shown in the form of flowchart in Figure 1)

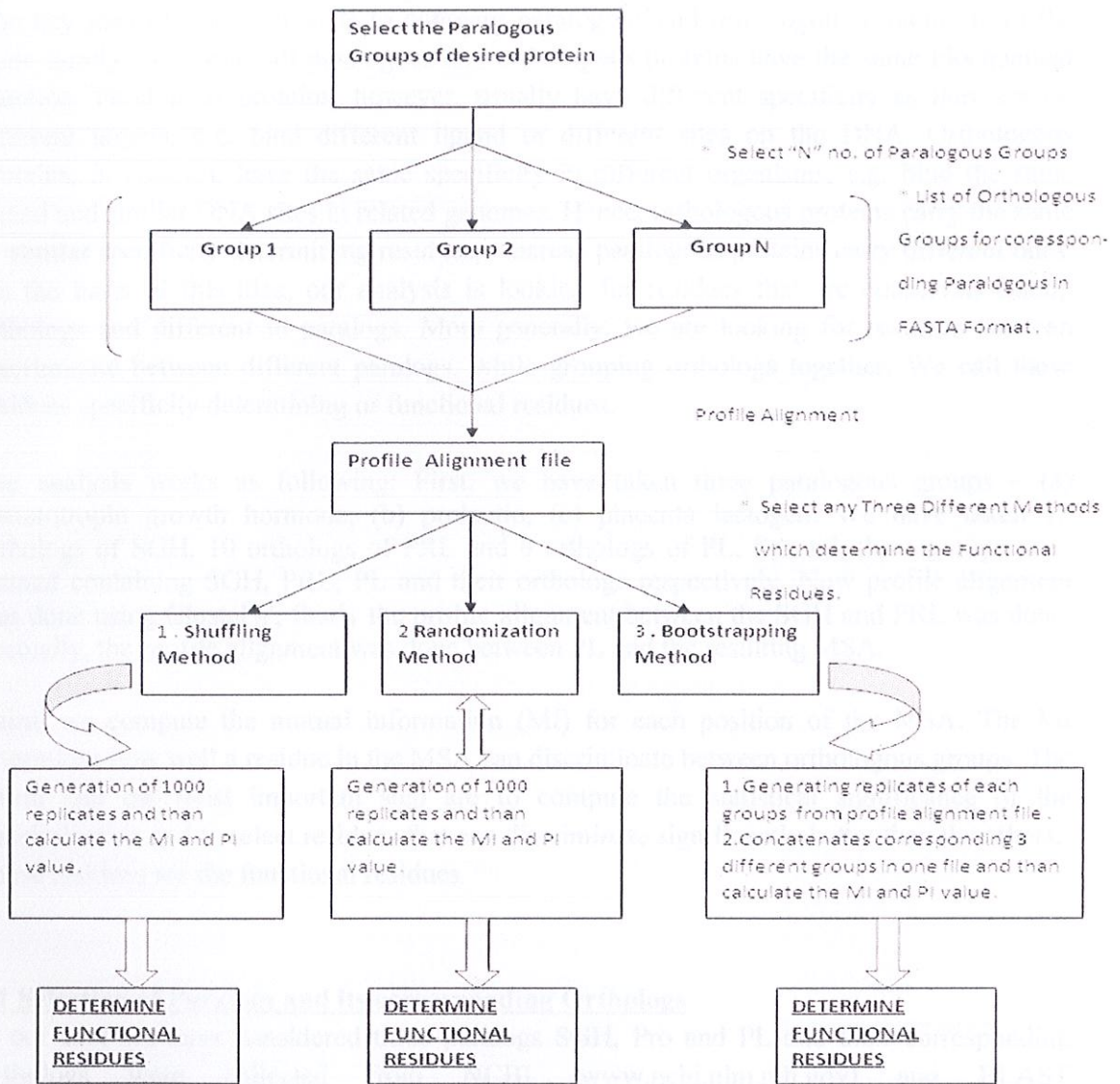


Figure 1: Steps followed in this project.



MATERIALS AND METHODS

The key idea of this method is to compare paralogous and orthologous proteins from the same family. As a rule, all paralogous and orthologous proteins have the same biochemical function. Paralogous proteins, however, usually have different specificity as they act on different targets, e.g. bind different ligand or different sites on the DNA. Orthologous proteins, in contrast, have the same specificity in different organisms, e.g. bind the same ligand and similar DNA sites in related genomes. Hence, orthologous proteins carry the same or similar specificity-determining residues, whereas paralogous proteins carry different ones. On the basis of this idea, our analysis is looking for residues that are conserved among orthologs and different in paralogs. More generally, we are looking for residues that can discriminate between different paralogs, while grouping orthologs together. We call these residues specificity determining or functional residues.

The analysis works as following: First, we have taken three paralogous groups – (a) somatotropin growth hormone, (b) prolactin, (c) placenta lactogen. We have taken 15 orthologs of SGH, 10 orthologs of PRL and 9 orthologs of PL. Second, three groups were formed containing SGH, PRL, PL and their orthologs respectively. Now profile alignment was done using ClustalW, firstly the profile alignment between the SGH and PRL was done. Secondly, the profile alignment was done between PL and the resulting MSA.

Third, we compute the mutual information (MI) for each position of the MSA. The MI determines how well a residue in the MSA can discriminate between orthologous groups. The fourth and the most important step are to compute the statistical significance of the discrimination and to select residues that can discriminate significantly better than the others. These residues are the functional residues.

3.1 Selection of Paralogs and its corresponding Orthologs

In our case we have considered three paralogs SGH, Pro and PL and their corresponding Orthologs were collected from NCBI (www.ncbi.nlm.nih.gov) and BLAST (www.ncbi.nlm.nih.gov/blast/Blast.cgi) databases and all the signal protein were removed with the aid of Swiss Pdb (<http://expasy.org/spdbv>). Then phylogenetic trees were constructed using the neighbor joining procedure implemented in PHYLIP.¹⁵

3.2 Profile Alignment

Now three groups were formed containing 15 Orthologs of SGH, 10 Orthologs of PRL and 9 Orthologs of PL. Now profile alignment was done using ClustalW, firstly the profile alignment between the SGH and Pro was done. Secondly, the profile alignment was done between PL and the resulting MSA. The final profile alignment file was saved in the format of .phy and .dnd extension which act as an input for PHYLIP to construct phylogenetic trees.

3.3 Mutual Information

To identify residues that can discriminate between paralogous proteins (different specificity), merging orthologs (same specificity) together we use the MI as a measure of association with the specificity. MI is frequently used in computational biology for co-variational analysis in RNA and proteins.

If $x = 1, \dots, 20$ is a residue type, $y = 1, \dots, Y$ is the specificity index which is the same for all proteins of the same specificity group and is different for different groups, and Y is the total number of specificity groups, then the mutual information at position i of the MSA is:

$$I_i = \sum_{x=1, \dots, 20} \sum_{y=1, \dots, Y} f_i(x, y) \log \frac{f_i(x, y)}{f_i(x) f_i(y)}$$

where $f_i(x)$ is the frequency of residue type x in position i of the MSA, $f_i(y)$ is the fraction of proteins belonging to the group y , and $f_i(x, y)$ is the frequency of residue type x in the group y at position i . MI has several important properties: (1) it is non-negative; (2) it equals zero if and only if x and y are statistically independent; and (3) a large value of I_i indicates a strong association between x and y . Unfortunately, a small sample size and a biased composition of each column in the MSA influences I_i a lot. For example, positions with less conserved residues tend to have higher MI. Hence, we cannot rely on the value of I_i as an indicator of specificity association, instead we estimate the statistical significance of I_i .

3.4 Statistical Significance

Since mutual information can be biased due to the small sample size or biased amino acid composition, we cannot rely on the value of mutual information to identify the specificity determinants. Instead, we compute the statistical significance $P(I)$ of the mutual information and use it together with I to predict the functional residues. Calculation of statistical significance is the most important component of the method. We followed three methods to calculate $P(I)$ they are shuffling, randomization and bootstrapping methods.

3.4.1 Method 1

We first compute I^{sh} using shuffling, then transform it to I^{exp} using the maximum likelihood estimator to take into account higher similarity between orthologs and finally compute the desired statistical significance $P(I)$.

We need to take into account the fact that orthologs are more closely related than paralogs. Due to this fact, sequence similarity between orthologs is higher than between paralogs. As a result, any position in the MSA has certain association with grouping by orthology. Specificity determinants, however, must have stronger association with this grouping than

any position on average. To compute $P(i)$ we start from a null-hypothesis that amino acid residues in all positions of the MSA have the same association with grouping by orthology.

Consider the MSA $a^m_i, i=1, \dots, L, m=1, \dots, M$, where a^m_i is the residue in position i of the m th protein, L is the length of the alignment (in our case L is 249) and M is total number of aligned proteins (in our case M is 34). For each position i we take a column a_i of the MSA and randomly shuffle this column. Next we compute the mutual information of the shuffled (a^{sh}_i) grouping: $I^{sh} = I(a^{sh})$. This procedure is repeated 10^4 times to get the distribution of the mutual information for a shuffled column.

To compute expected MI I^{exp} we make transformation:

$$I^{exp}_i = \alpha I^{sh}_i + \beta$$

Here α and β are calculated by the formula:

$$\frac{\langle I^{sh}_i \rangle}{\sigma(I^{sh}_i)} = A \frac{I_i}{\sigma(I^{sh}_i)} + B \frac{1}{\sigma(I^{sh}_i)}$$

And the obtain $\alpha=1/A$ and $\beta=-1/B$. Here $\langle I^{sh}_i \rangle$ and $\sigma(I^{sh}_i)$ are the mean and the variance of I^{sh}_i obtained by 10^4 random shufflings.

After α and β are computed, we obtained the desired probability:

$$P_i = P(I_i) = \frac{1}{\sqrt{2\pi\sigma(I_i^{exp})}} \int_{I_i}^{\infty} \exp\left(-\frac{(t - \langle I_i^{exp} \rangle)^2}{2\sigma^2(I_i^{exp})}\right) dt$$

$$= \frac{1}{\sqrt{2\pi}} \int_{Z_i}^{\infty} \exp\left(-\frac{z^2}{2}\right) dz$$

where:

$$Z_i = \frac{I_i - \langle I_i^{exp} \rangle}{\sigma(I_i^{exp})}$$

Very low P_i indicates that the null-hypothesis does not hold for position i and residues in this position are in fact stronger associated with the specificity grouping than the whole proteins. Thus positions in the MSA that exhibit low P_i and high I_i are the Functional residues.

3.4.2 Method 2

This model does not utilize shuffling to compute I^{exp} , Instead we model evolution of the protein family and generate a set of pseudo-random protein sequences Using obtained pseudo-random proteins we compute mutual information I^{md} . Finally, we compute $P_i = P(I_i)$ as the probability of observing mutual information above I_i for the pseudo-random proteins.

We start from a null-hypothesis that all positions in the MSA have the same association with specificity grouping. To compute P_i ; we need to generate sequences that have the same intra-group and inter-group similarity as the orthologs and the paralogs, respectively. This is achieved by simulating evolution of these proteins in the following manner:

- (i) Generate a “parent” sequence $b^y = b^y_i, i=1 \dots\dots L$ (in our case L is 249) for each group of orthologs $y=1, \dots\dots Y$ (in our case Y is 3). An amino acid residue b^y_i is generated randomly from the distribution of amino acids at position i $f_i(x)$. This step simulates evolution of paralogous proteins by duplication.
- (ii) Generate a sequence of the m th protein c^m_i from the “parent” sequence of its group y . We assume that during speciation, that followed duplication, some amino acids did not get substituted. We simulate this by introducing the probability μ (set $\mu=0.85$) of inheriting an amino acid from the “parent” protein without a substitution. Hence $c^m_i = b^y_i$ with probability μ , and c^m_i is taken randomly from $f_i(x)$ with probability $1 - \mu$. This step simulates evolution of orthologs through speciation.

After pseudo-random correlated sequences are generated, we compute I^{md}_i . The sequences (including “the parents”) are generated in 10^3 independent runs yielding the distribution of the mutual information $f_i(I^{md})$. Assuming normal distribution of I^{md} we get P_i as:

$$P_i = P(I_i) = \frac{1}{\sqrt{2\pi}\sigma(I_i^{md})} \int_{I_i}^{+\infty} \exp\left(-\frac{(I - (I_i^{md}))^2}{2\sigma^2(I_i^{md})}\right) dI$$

$$= \frac{1}{\sqrt{2\pi}} \int_{Z_i}^{+\infty} \exp\left(-\frac{z^2}{2}\right) dz$$

where:

$$Z_i = \frac{I_i - (I_i^{md})}{\sigma(I_i^{md})}$$

Very low P_i indicates that the null-hypothesis does not hold for position i and residues in this position are in fact stronger associated with the specificity grouping than the whole proteins. Thus positions in the MSA that exhibit low P_i and high I_i are the Functional residues.

3.4.3 Method 3

In this method MI is calculated by parametric bootstrapping, in this case replicates of different groups are generated using tool named pseq-gen which runs in Linux and takes length of the alignment and number of replicates to be generated as an input. In our case length of the alignment was 249 and number of replicates was 100. Now all the replicates from three groups were taken together and therefore we had 100 files.

MI was then calculated which act as an input for calculating P_i as:

$$P_i = P(I_i) = \frac{1}{\sqrt{2\pi}\sigma(I_i^{mid})} \int_{I_i}^{+\infty} \exp\left(-\frac{(I - (I_i^{mid}))^2}{2\sigma^2(I_i^{mid})}\right) dI$$
$$= \frac{1}{\sqrt{2\pi}} \int_{Z_i}^{+\infty} \exp\left(-\frac{z^2}{2}\right) dz$$

where:

$$Z_i = \frac{I_i - (I_i^{mid})}{\sigma(I_i^{mid})}$$

Very low P_i indicates that the null-hypothesis does not hold for position i and residues in this position are in fact stronger associated with the specificity grouping than the whole proteins. Thus positions in the MSA that exhibit low P_i and high I_i are the functional residues.

RESULTS AND DISCUSSION

We have chosen Somatotropin Growth Hormone orthologs for our analysis because (1) available experimental and structural information¹⁶ can be used to assess our predictions (2) large amount of information is known (3) its function is known and (4) the relative stability of their respective 1:2 complexes.

Figure 2. Presents the output of MSA of the individual orthologous groups i.e. (A) SGH, (B) PL and (C) PRL done through ClustalW.

(A)

```

-----VEP-----ISLYNLFSAVNRAQHLHLLAAEITYKEFERSSIPPEA---HRQLSKTSPLAGCYSDSIPTPT
-----FESQRLFNNAVIRVQHLHQLAAKMMDDFEALLPBE---RKQLSKIFPLSFCNSDSIEAPA
-----MENQRLFNIAVNRVQHLHLLAQKMFNDFEGTLLPDE---RRQLNKIFLLDFCNSDSIVSEPI
-----IENQRLFNIAVSRVQHLHLLAQKMFNDFDGTLLPDE---RRQLNKIFLLDFCNSDSIVSEPV
-----IENQRLFNIAVSRVQHLHLLAQKMFNDFDGTLLPDE---RRQLNKIFLLDFCNSDSIVSEPV
-----IENQRLFNIAVSRVQHLHLLAQKMFNDFDGTLLPDE---RRQLNKIFLLDFCNSDSIVSEPV
-----MENQRLFNIAVNRVQHLHLLAQKMFNDFEGTLLSDE---RRQLNKIFLLDFCNSDSIVSEPI
-----MENQRLFNIAVNRVQHLHLLAQKMFNDFEGTLLPDE---RRQLNKIFLLDFCNSDSIVSEPI
-----MENQRLFNIAVNRVQHLHLLAQKMFNDFEGTLLPDE---RRQLNKIFLLDFCNSDSIVSEPI
-----MENQRLFNIAVNRVQHLHLLAQKMFNDFEGTLLPDE---RRQLNKIFLLDFCNSDSIVSEPI
-----QPI-----TENQRLFSIAVGRVQYHLVAKKLFSDFENSLQLED---QRLLNKIASREFCHSDNFLSEPI
-----QPI-----PNNQHLFSMAVSRIRHLLHRAQRLFANFESSLQSD---QRLLNKIFLQDFCNSDYIISPI
-----QPM-----TDSQR-FSIAVSRIRYHLQVAQRSFSSSLSAED---QRLLNKIFLQDFCNSDYIRSEPI
-----QPI-----TDSQRLFSIAVSRVQHLHLLAQRRFSEFESSLQTEE---QRLLNKIFLQDFCNSDYIISPI
-----QOI-----TDSQRLFSIAVNRVTHLHLLAQRLFSDFESSLQTEE---QRLLNKIFLQDFCNSDYIISPI
-----QOI-----TDSQRLFSIAVNRVTHLYLLAQRLFSDFESSLQTEE---QRLLNKIFLQDFCNSDYIISPI
    
```

(B)

```

-----MAAGSRTSLLLAFAALLCLPWLOEAG-AVQTVPLSRLFDHAMLQAHRAHQLAIDTYQEFEEAYIL
-----MAPGSRTSLLLAFAGLLCLPWLOEAG-AFPTIPLSRLFDHAMLQAHRAHQLAIDTYQEFEEAYIP
-----MAAGSRTSLLLAFAALLCLPWLOEAG-AVQTVPLSRLFKEAMLQAHRAHQLAIDTYQEFEEAYIP
-----MAPGSRTSLLLAFAALLCLPWLOEAG-AIQTVPLSRLFDHAMLQAHRAHQLAIDTYQEFEEAYIP
MQLTTLTSGSGMQLLLLVSSLLL-WENVASKPTAIVSTDDLYHRLVESHNTFIMAADVYREFDINFAPK
MQLTTLTSGSGMQLLLLVSSLLL-WENVASKPTAIVSTDDLYHRLVESHNTFIMAADVYREFDINFAPK
MQLTTLTSGSGMQLLLLVSSLFL-WENVSSKPTAMVPTEDLYTRLAELSHNTFILAADVYREFDLDFD
MQLTTLTSGSGMQLLLLVSSLLL-WENVSSKPTAMVPTDDLYTRLAELSHNTFILAADVYREFDLDFD
MQLTTLTSGSGMQLLLLVSSLLL-WENVSSKPTAMVPTDDLYTRLAELSHNTFILAADVYREFDLDFD
    
```

(B)

```

-----LPICPIGSVNCQVSLGELFDRAVKLSHYIHYLSSEIFNEFDERYAQGR---GFITRAV---
-----LPICPGGAARCOVTLRDLFDRAVLSHYIHNLSSSEMFSEFDKRYTHGR---GFITKAI---
-----LPICPSGAVNCQVSLRDLFDRAVILSHYIHNLSSSEMFNEFDKRYAQGR---GFITKAI---
-----LPICPSGAVNCQVSLRELFDRAVILSHYIHNLSSSEMFNEFDKRYAQGR---GFVTKAI---
-----IPISDLDRASQRSDTLHSLSTTLTQDLDSEHPPPMG---RVITPRP---
-----IGLSDLMERASQRSKLSLSTSLTKDLDSEHPPPMG---RVMMPRP---
-----VNLNDLLDRASQSDKMSLSTSLTNDLDSEHSSVG---GKLM-RP---
-----VGLNDLLERASQSDKLSLSTSLTNDLDSEHPPPVG---RVMMPRP---
-----VGLNDLLERASQSDKLSLSTSLTNDLDSEHPPPVG---RVMMPRP---
-----VGLNDLLERASQSDKLSLSTSLTNDLDSEHPPPVG---RVMMPRP---
    
```

Now profile alignment was done using ClustalW, firstly the profile alignment between the SGH and Pro was done. Secondly, the profile alignment was done between PL and the resulting MSA. Figure 3 presents the final profile alignment between the three orthologous groups.

```

-----VEP-----ISLYNLFTSAVNRAQHLHTLAAEYKEFERSIPPEA---HRQLSKTSPLAGCYSDSIPTPTGKDETQEKSDGYLLRISSALIQ
-----FESQRLFNNAVIRVQHLHLLAQKMFNDFEGTLLPDE---RRQLNKIFLLDFCNSDSIVSPIDKHETQKSSVLKLLHISFRLIIE
-----MENQRLFNIAVNRVQHLHLLAQKMFNDFEGTLLPDE---RRQLNKIFLLDFCNSDSIVSPIDKHETQKSSVLKLLHISFRLIIE
-----IENQRLFNIAVSRVQHLHLLAQKMFNDFEGTLLPDE---RRQLNKIFLLDFCNSDSIVSPVKKHETQKSSVLKLLHISFRLIIE
-----IENQRLFNIAVSRVQHLHLLAQKMFNDFEGTLLPDE---RRQLNKIFLLDFCNSDSIVSPVKKHETQKSSVLKLLHISFRLIIE
-----IENQRLFNIAVSRVQHLHLLAQKMFNDFEGTLLPDE---RRQLNKIFLLDFCNSDSIVSPVKKHETQKSSVLKLLHISFRLIIE
-----MENQRLFNIAVNRVQHLHLLAQKMFNDFEGTLLSDE---RRQLNKIFLLDFCNSDSIVSPIDKHETQKSSVLKLLHISFRLIIE
-----MENQRLFNIAVNRVQHLHLLAQKMFNDFEGTLLPDE---RRQLNKIFLLDFCNSDSIVSPIDKHETQKSSVLKLLHISFRLIIE
-----MENQRLFNIAVNRVQHLHLLAQKMFNDFEGTLLPDE---RRQLNKIFLLDFCNSDSIVSPIDKHETQKSSVLKLLHISFRLIIE
-----MENQRLFNIAVNRVQHLHLLAQKMFNDFEGTLLPDE---RRQLNKIFLLDFCNSDSIVSPIDKHETQKSSVLKLLHISFRLIIE
-----QPI-----TENQRLFSIAVGRVQYLHLVAKKLFSDFENSQLED---QRLLNKIASKEFCSDNFIKSPIDKHETQSSVQKLLSVYRLLIIE
-----QPI-----PNNQHLSMAVSRVQHLHLLAQKMFNDFEGTLLSDD---QRLLNKIFLQDFCNSDYIISPIDKHETQSSVQKLLSVYRLLIIE
-----QPM-----TDSQR--FSIAVSRVQHLHLLAQKMFNDFEGTLLSDE---QRLLNKIFLQDFCNSDYIISPIDKHETQSSVQKLLSVYRLLIIE
-----QPI-----TDSQR--FSIAVSRVQHLHLLAQKMFNDFEGTLLSDE---QRLLNKIFLQDFCNSDYIISPIDKHETQSSVQKLLSVYRLLIIE
-----QQI-----TDSQR--FSIAVSRVQHLHLLAQKMFNDFEGTLLSDE---QRLLNKIFLQDFCNSDYIISPIDKHETQSSVQKLLSVYRLLIIE
-----QQI-----TDSQR--FSIAVSRVQHLHLLAQKMFNDFEGTLLSDE---QRLLNKIFLQDFCNSDYIISPIDKHETQSSVQKLLSVYRLLIIE
-----LPICPIGAVNCQVSLGELFDRVAVLSHYIHLSEIFNEFDERYAQGR---GFITKAV---NGCHTSSLTPEDKEQAQIHHEDLLNLVVGVLFI
-----LPICPGGAARCVTLRDLFDRVAVLSHYIHLSEIFNEFDERYAQGR---GFITKAV---NSCHTSSLTPEDKEQAQIHHEDLLNLVVGVLFI
-----LPICPSGAVNCQVSLRDLFDRVAVLSHYIHLSEIFNEFDERYAQGR---GFITKAV---NSCHTSSLTPEDKEQAQIHHEDLLNLVVGVLFI
-----LPICPSGAVNCQVSLRDLFDRVAVLSHYIHLSEIFNEFDERYAQGR---GFVTKAI---NSCHTSSLTPEDKEQAQIHHEDLLNLVVGVLFI
-----IPISDLDRASQSDTLHSLSTSLTNDLDSHFPPMG---RVITPRP---SMCHTSSLQIPNDKQDALKVPEDLLSLARSLLIIE
-----IGLSDLMERASQSDTLHSLSTSLTNDLDSHFPPMG---RVMMPRP---SMCHTSSLQIPNDKQDALKVPEDLLSLARSLLIIE
-----VNLNDLLERASQSDTLHSLSTSLTNDLDSHFPPMG---GKLM--RP---SMCHTSSLQIPNDKQDALKVPEDLLSLARSLLIIE
-----VGLNDLLERASQSDTLHSLSTSLTNDLDSHFPPMG---RVMMPRP---SMCHTSSLQIPNDKQDALKVPEDLLSLARSLLIIE
-----VGLNDLLERASQSDTLHSLSTSLTNDLDSHFPPMG---RVMMPRP---SMCHTSSLQIPNDKQDALKVPEDLLSLARSLLIIE
-----VGLNDLLERASQSDTLHSLSTSLTNDLDSHFPPMG---RVMMPRP---SMCHTSSLQIPNDKQDALKVPEDLLSLARSLLIIE
-----VGLNDLLERASQSDTLHSLSTSLTNDLDSHFPPMG---RVMMPRP---SMCHTSSLQIPNDKQDALKVPEDLLSLARSLLIIE
-----MAAGSRTSLLLAFALLCLPWLQEG--AVQTVPLSRLFDHAMLQAHRAHQLAIDTYQEFEEAYIPKDQKYSFLHDSQ---TSFCFSDSIPTPSNMEETQKSNLELLRISLLIIE
-----MAPGRTSLLLAFALLCLPWLQEG--AFPTIPLSRLFDHAMLQAHRAHQLAIDTYQEFEEAYIPKDQKYSFLHDSQ---TSFCFSDSIPTPSNMEETQKSNLELLRISLLIIE
-----MAAGSRTSLLLAFALLCLPWLQEG--AVQTVPLSRLFDHAMLQAHRAHQLAIDTYQEFEEAYIPKDQKYSFLHDSQ---TSFCFSDSIPTPSNMEETQKSNLELLRISLLIIE
-----MAPGRTSLLLAFALLCLPWLQEG--AIQTVPLSRLFDHAMLQAHRAHQLAIDTYQEFEEAYIPKDQKYSFLHDSQ---TSFCFSDSIPTPSNMEETQKSNLELLRISLLIIE
MQLTLTLSGSGMQLLLLVSSLLL--WENVASKPTAIVSTDDLYHRLVEQSHNTFIMAADVYREFDINFAKR---SMMKDR---LPLCHTASIHTPENLEEVHEMKTEDFLNSIINVSI
MQLTLTLSGSGMQLLLLVSSLLL--WENVASKPTAIVSTDDLYHRLVEQSHNTFIMAADVYREFDINFAKR---SMMKDR---LPLCHTASIHTPENLEEVHEMKTEDFLNSIINVSI
MQLTLTLSGSGMQLLLLVSSLLL--WENVSSKPTAMVPTDDLYTRLAELSHNTFILAADVYREFDLDFFDK---TWITDRT---LPLCHTASIHTPENREEVHEIKTEDLLKAMINVSII
MQLTLTLSGSGMQLLLLVSSLLL--WENVSSKPTAMVPTDDLYTRLAELSHNTFILAADVYREFDLDFFDK---TWITDRT---LPLCHTASIHTPENREEVHEIKTEDLLKAMINVSII

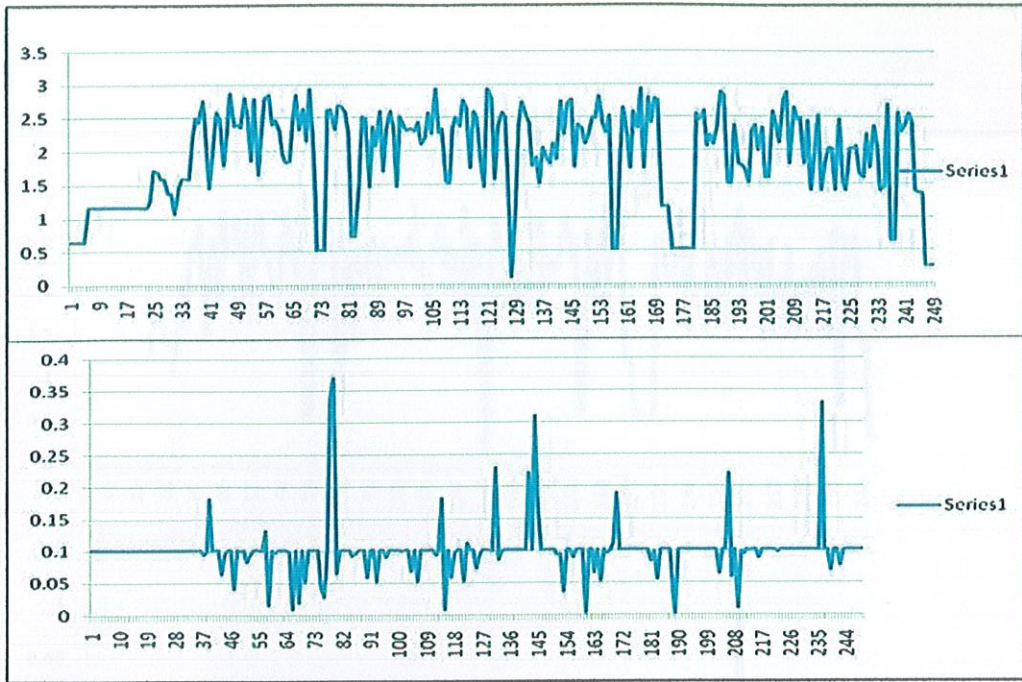
```

Figure 3: Profile alignment between three paralogs groups SGH, Pro, and PL

Figure 4 presents the mutual information I_i and the probability $P(I_i)$ computed for the three orthologous groups taken together using all the three methods. This plot reveals that very few positions have both low $P(I_i)$ and high I_i . Amino acid residues in these positions have strong association with functional grouping (stronger than sequences on average), indicating the role of these positions in determining different specificities of different groups of orthologs.

(A)

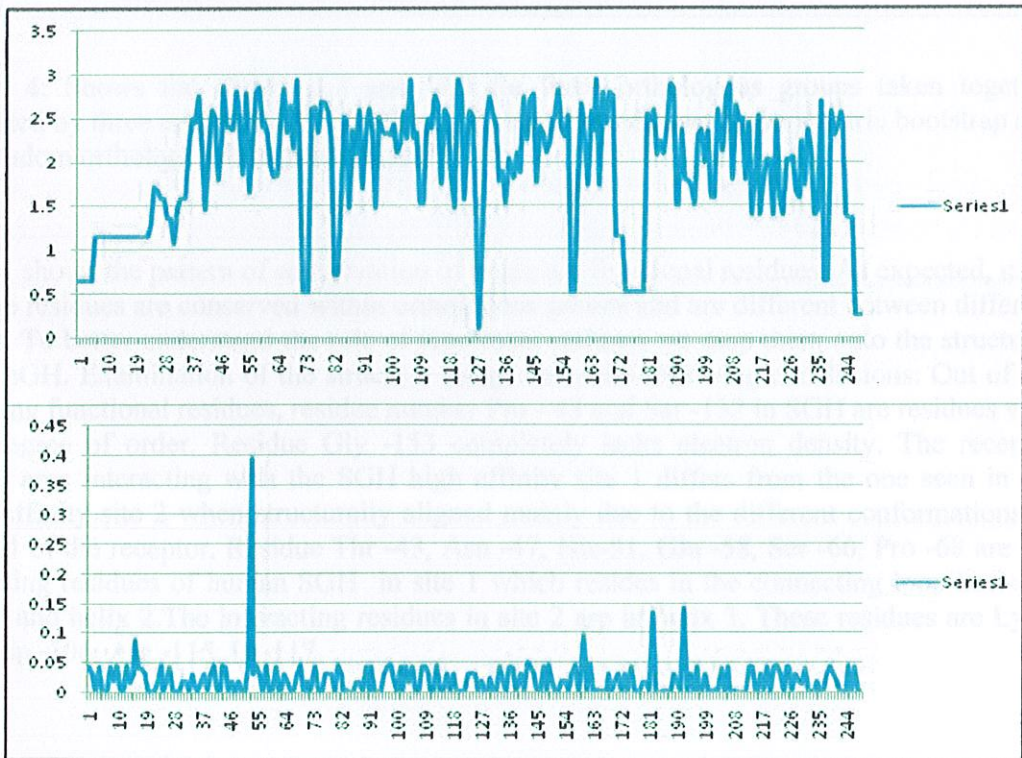
I_i



(B)

I_i

$P(I_i)$



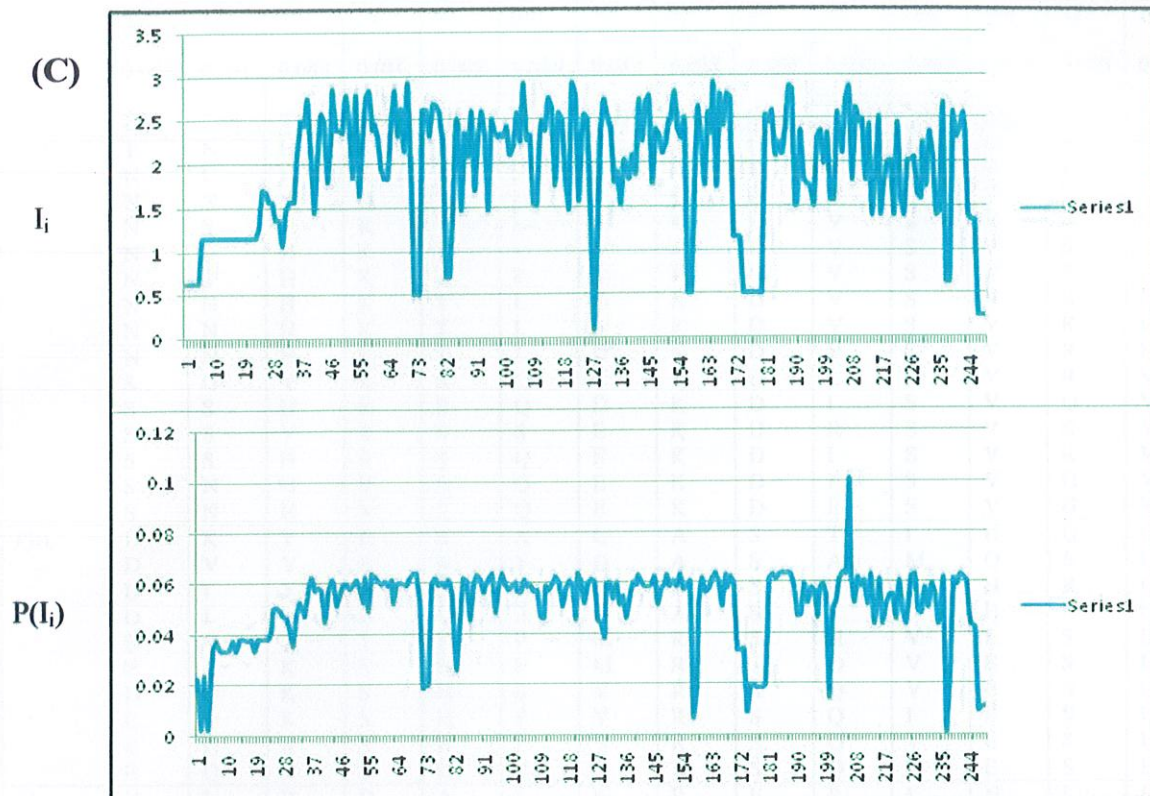


Figure 4: Shows the observed I and $P(I)$ for three orthologous groups taken together calculated by three methods: (A) Shuffling, (B) Replicate dataset by parametric bootstrap and (C) Random orthologs group generation.

Table 1 shows the pattern of conservation of predicted functional residues. As expected, most of these residues are conserved within orthologous groups and are different between different groups. To better understand the role of functional residues we map them onto the structures of the hGH. Examination of the structure brings us to the following conclusions: Out of the following functional residues, residue number Pro - 43 and Ser -132 in SGH are residues with poor degree of order. Residue Gly -153 completely lacks electron density. The receptor surface area interacting with the SGH high affinity site 1 differs from the one seen in the lower affinity site 2 when structurally aligned mainly due to the different conformations of Trp-104 of the receptor. Residue Thr -43, Asn -47, His-51, Glu -58, Ser -66, Pro -68 are the interacting residues of human SGH in site 1 which resides in the connecting loop between helix 1 and helix 2. The interacting residues in site 2 are in helix 3. These residues are Lys -104, Asp -106, Ala -115, Ile-117.

Table 1 : Showing Functional Residues in three orthologous groups (SGH, PRL and PL)

Residues	43	47	51	58	66	68	70	80	90	93	104	106	115	117	121
P(I)	0.064	0.041	0.083	0.016	0.009	0.019	0.052	0.065	0.058	0.052	0.067	0.052	0.008	0.059	0.052
(I)	2.58	2.86	2.79	2.84	2.83	2.63	2.98	2.58	2.58	2.59	2.57	2.92	2.64	2.58	2.92
SGH	T N N N N N N N N S S S S S	N I N S S N N N N G S S S N	H H H H H H H H Y H Y H H H	E K K K K K K K K R R R R R	S A T T T T T T S S S S S	P L L L L L L L Q Q Q Q Q	E E D D D D D D E E E E E	K K K K K K K K K K K K K K K	D D D D D D D D D D D D D D D	P E V V V V V V L R I I I	K S S S S S S S S S S S S S	D V V V V V V V V V V V V V	A R R R R R R R R R R R R R	I I I I I I I I I I I I I I	V E E E E E E E E E E E E E
PRL	D D D D D D E E E	K V I I Q Q Q Q Q	Y Y Y T K K K K K	E E E T S S S S S	R R R H H H H H H	A T A P P S P P	G G G M M V V V	A A A R R R R R	S S S S S S S S S	T A P S H Q Q Q Q	I M I I V V I V V	H Q H H E E E E E	G S R R S S S S S	L L L L L L L L L	N N N N N N S S S
PL	D D K D H H T T T	L L L E E E E E	R P R N N S N N	D D D D D D D D D	A A A N N D D D	I I I A A F F F	K K K R R K K K	P S S R R R R R	E D D A A A A A	P P P H H H H H	K K K M M I T I	N N N N T T T T	L L L L N N N N	I I I S S S S	L L L L K K K K

(Continued)

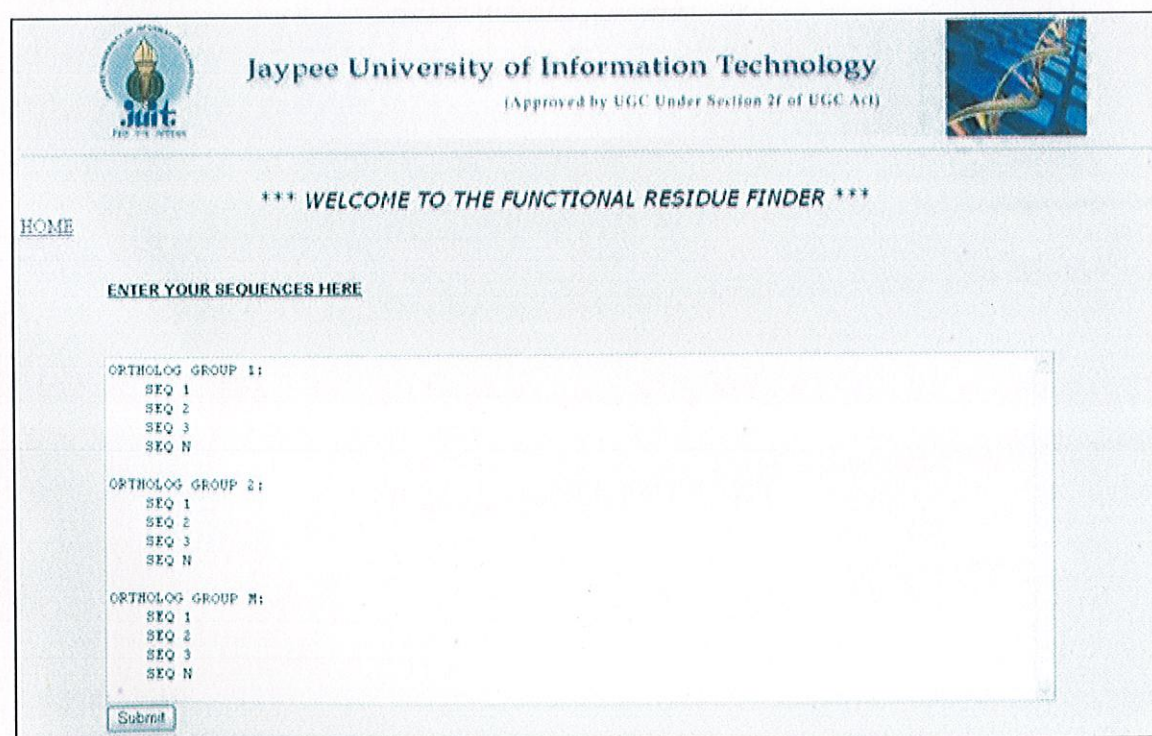
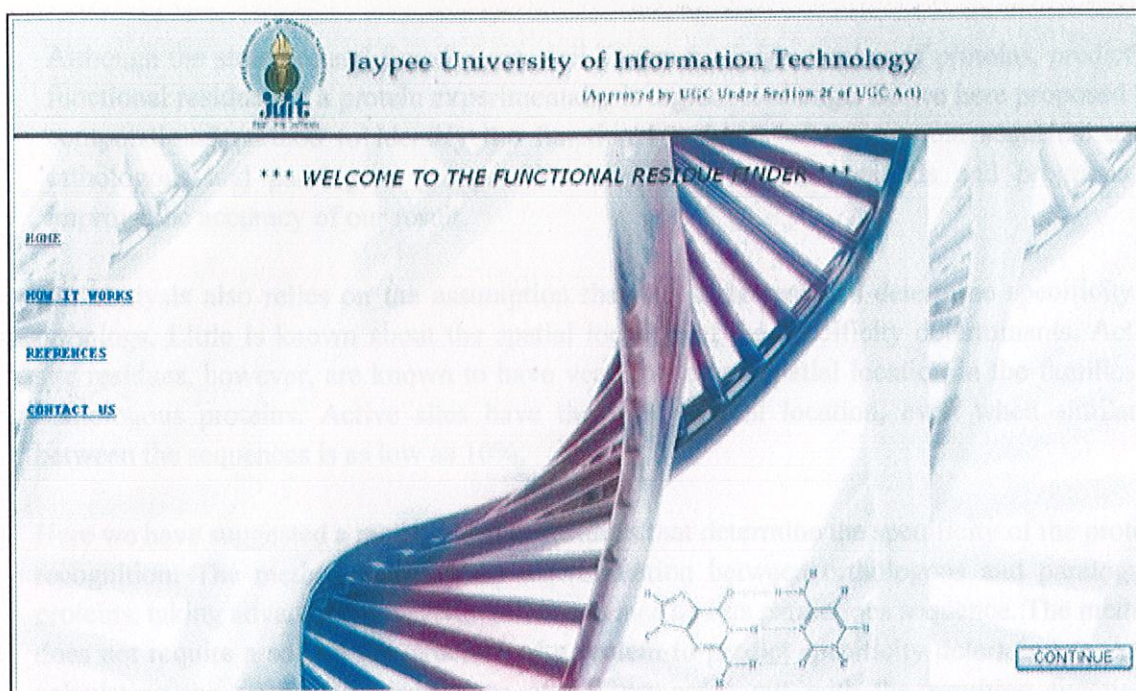
(Here I is mutual information and P(I) is statistical significance of mutual information)

Table 1: (Continued)

Residues	125	132	153	160	163	165	181	183	189	203	207	209	216	239	242
P(I)	0.072	0.086	0.030	0.002	0.065	0.052	0.083	0.054	2*10 ⁻⁵	0.06	0.058	0.01	0.087	0.068	0.076
(I)	2.566	2.532	2.815	2.645	2.580	2.928	2.540	2.595	2.88	2.58	2.88	2.63	2.52	2.573	2.558
SGH	K R Q Q Q Q Q R R R R R	S - S S S S S S G G G G	G G G G G G G G G G G G	M I I I I I I I I M I I	V C N S S S S N N N N N	D D E D D D D D E D D D	I L L L L L L L F F F Y Y	D E D D D D D D E D D D D	- - L L L L L L L L L H H	R - G G G G G - G G G G	G N N N N N N N - L P S S	M R R R R R R - R R R R	A S A A A A A A A A A A	F S S S S S S F S S S S	S S A A A A A A A A A A
PRL	I Y Y Y V L A A A A	V V V A A A A A	Q Q Q H Y H N N N	L L L G G G V G	M M M L L L L L	K L K I I R H H R	H H Q G G G G G	G E G R A S S S S	I I V V S S S S S	D D D D D D D D D	R R R R R R R R	F S F F T I V V V	H H H S S S S S	I I V V L M K K K	S N S S E E E E E
PL	R R R R K K K K	M M M A A A A A	Y Y Y K K R R R	E E E L L L L L	I I I L L L L L	T T T T T T T T	L L L V V I I I	Q Q Q E E E E E	D D D D D D D D	H H H D D D N N	A A A E H H H H	L L L F F F F F	Y - H H R R R R	V - V V V V V V V	G - G N N N N N

(Here I is mutual information and P(I) is statistical significance of mutual information)

Figure 5: Screen shot of our tool.



CONCLUSION

Although the structure and function are well known for large number of proteins, predicting functional residues of a protein experimentally is a great challenge. So we here proposed the computational method to identify the functional residues of any protein sequence using orthologous and paralogous groups. We have used various methods and programs to improve the accuracy of our result.

Our analysis also relies on the assumption that the same residues determine specificity of orthologs. Little is known about the spatial location of the specificity determinants. Active site residues, however, are known to have very conserved spatial location in the families of homologous proteins. Active sites have the same spatial location, even when similarity between the sequences is as low as 10%.

Here we have suggested a method to find residues that determine the specificity of the protein recognition. The method is based on discrimination between orthologous and paralogous proteins, taking advantage of several SGH sequence and its paralogous sequence. The method does not require a solved 3D structure of a protein to predict specificity determinants. After calculating the statistical significance of MI we came out with the resulting functional residues. When we crosschecked our results with the given literature we found that the results were very satisfactory.

FUTURE PROSPECTIVE

The future prospective of this tool would be increasing the accuracy and efficiency. This would be done by adding sequence weights and testing this tool with more number of orthologs. We are planning to put the tool online so that it can be further used by other people for their research work.

1. Wollenberg, Kurt & Ashley W. (2000). Separation of phylogenetic and functional associations in biological sequences - by using the parametric bootstrap. *PNAS* 97, 3288-3291.
2. Fitch, W. (1970). Distinguishing homologous from analogous proteins. *Syst. Zool.* 19, 99-113.
3. Fitch, W. (2000). Nomenclature: a personal view on some of the problems. *Trends Genet.* 16, 237-241.
4. Drenth, A. (1999). Structures and Mechanisms in Protein Science: A Guide to Enzyme Catalysis and Protein Folding. WH Freeman & Co, San Francisco.
5. Ballinger, M., Tom, J. & Wells, J. (1996). Fertilisin: a variant of subtilisin type II engineered for cleaving tribock substrates. *Biochemistry*, 35, 13579-13585.
6. DeVlaming, V. (1979) Actions of prolactin among the vertebrates. In *Hormones and evolution* (ed., Barrington, E.J.W.), 561-592.
7. Saha, Y.N. (1995) Structural variants of prolactin: occurrence and physiological significance. *Endocrine Rev.* 16, 354-369.
8. Griffin, V., Shiverick, K.T., Kelly, P.A. & Martial, J.A. (1996). Sequence-function relationships within the expanding family of prolactin, growth hormone, placental lactogen, and related proteins in mammals. *Endocrine Rev.* 17, 385-410.
9. Nicoll, C.N., Mayer, G.C. & Russell, S.M. (1986). Structural features of prolactins and growth hormones that can be related to their biological properties. *Endocrine Rev.* 7, 193-203 (1986).
10. Genter, A., Grakalovic, J., Losburger, C.F., Wu, S. & Djinovic, J. (1996). Real-time kinetic measurements of the interactions between lactogenic hormones and prolactin receptor extracellular domains from several species support the model of hormone-induced transient receptor dimerization. *J. Biol. Chem.* 271, 29482-29491.

BIBLIOGRAPHY

Research papers

1. Mirny, L., & Gelfand, M. (2002). Using Orthologous and Paralogous Proteins to Identify Specificity-determining Residues in Bacterial Transcription Factors. *J. Mol. Biol.* 321, 7-20.
2. Wollenberg, Kurt & Atchley W. (2000). Separation of phylogenetic and functional associations in biological sequences by using the parametric bootstrap. *PNAS* 97, 3288-3291
3. Fitch, W. (1970). Distinguishing homologous from analogous proteins. *Syst. Zool.* 19, 99-113.
4. Fitch, W. (2000). Homology a personal view on some of the problems. *Trends Genet.* 16, 227-231
5. Fersht, A. (1999). *Structure and Mechanism in Protein Science: A Guide to Enzyme Catalysis and Protein Folding*, WH Freeman & Co, San Francisco.
6. Ballinger, M., Tom, J. & Wells, J. (1996). Furilisin: a variant of subtilisin bpn0 engineered for cleaving tribasic substrates. *Biochemistry*, 35, 13579-13585.
7. DeVlaming, V. (1979) Actions of prolactin among the vertebrates. In *Hormones and evolution* (ed., Barrington, E.J.W.), 561-642
8. Sinha, Y.N. (1995) Structural variants of prolactin: occurrence and physiological significance. *Endocrine Rev.* 16, 354-369
9. Goffin, V., Shiverick, K.T., Kelly, P.A. & Martial, J.A. (1996). Sequence-function relationships within the expanding family of prolactin, growth hormone, placental lactogen, and related proteins in mammals. *Endocrine Rev.* 17, 385-410 .
10. Nicoll, C.S., Mayer, G.L. & Russel, S.M. (1986). Structural features of prolactins and growth hormones that can be related to their biological properties. *Endocrine Rev.* 7, 169-203 (1986).
11. Gertler, A., Grosclaude, J., Strasburger, C.J., Nir, S. & Djiane, J. (1996). Real-time kinetic measurements of the interactions between lactogenic hormones and prolactin-receptor extracellular domains from several species support the model of hormone-induced transient receptor dimerization. *J. Biol. Chem.* 271, 24482-24491.

12. De Vos, A.M., Ultsch, M. & Kossiakoff, A.A. (1992). Human growth hormone and extracellular domain of its receptor: crystal structure of the complex. *Science* **255**, 306–312.
13. Kelly, P.A., *et al.* (1991). The growth hormone/prolactin receptor gene family. *Oxford Surveys On Eukaryotic Genes* **7**, 29–50 (1991).
14. Thompson, J., Higgins, D. & Gibson, T. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucl. Acids Res.* **22**, 4673–4680.
15. Tuimala, J. (2006) A primer to phylogenetic analysis using Phylip Package. Fifth Edition.
16. Sundstrom, M., Lundqvist, T., Rodin, J., Giebel, L., Milligan, D., & Norsted, G. (1996) Crystal Structure of an Antagonist Mutant of Human Growth Hormone, G120R, in Complex with Its Receptor at 2.9 Å Resolution. *271*, 32197-32203

Websites

1. NCBI : www.ncbi.nlm.nih.gov
2. BLAST : www.ncbi.nlm.nih.gov/blast/Blast.cgi
3. SWISS PDB : www.expasy.org/spdbv
4. KEGG : www.genome.jp/kegg
5. RCSB : www.rcsb.org