

# **Bioinformatics Literacy in Pakistan:**

## **A survey based approach**

*Supervisor:* Dr. Habib Bukhari

*Researchers:* Abdullah Ahmed, Mariya Naeem, Sumaira Idrees

*Edited by* Ayesha Saeed Khan

# Table of Contents

Abstract.....	3
Summary .....	3
Introduction.....	5
The Gap between Biological and Computer Sciences: .....	6
The situation so far: .....	6
Materials and Methods: .....	8
Questionnaire.....	8
Section 1.....	8
Section 2.....	9
Section 3.....	9
Section 4.....	10
Results:.....	11
Conclusions: .....	19
Bioinformatics Literacy rates:.....	19
The tools used by students and professionals:.....	19
Reasons for low literacy:.....	21
Appendix A: URLs of Bioinformatics Tools.....	22
References: .....	23

## **Abstract**

This is a survey based approach to determine the level of Bioinformatics Literacy in Pakistan with Islamabad as a representative sample. It was conducted in academic institutions and research organizations in the Islamabad region. The aim of this study was threefold, to determine the level of Bioinformatics literacy amongst students, researchers, professionals and organizations in Pakistan, to determine the tools used and to determine the reasons, if any, for low Bioinformatics literacy levels. The respondents were tested using a questionnaire and a 5 point Lickert scale was used to interpret the results.

## **Summary**

In developed countries around the world Bioinformatics knowledge has already been integrated into university curriculum at both graduate and undergraduate level. Unfortunately this is not the case in most developing countries including Pakistan. Therefore there is a need to determine the degree and the extent of Bioinformatics literacy present in our country. Before we begin the teaching and developing our work force it is necessary to determine how Bioinformatics is used by researchers and students in their research and studies. This survey satisfies that very need.

This is a survey based approach to determine the level of Bioinformatics Literacy in Pakistan with Islamabad as a representative sample. The interviewees taking part in the survey were students, teachers and researchers from academic institutions and research organizations within Islamabad. The aim this study was threefold:

1. To determine the level of Bioinformatics literacy amongst students, researchers, professionals and organizations in Pakistan.
2. To determine which Bioinformatics tools (with emphasis towards the field of Microbiology) are currently being used by various students, researchers, professionals and organizations in Pakistan.
3. To determine the reasons if any for low Bioinformatics literacy in Pakistan.

Firstly a questionnaire was developed. The questionnaire was designed to obtain demographics data, to assess knowledge of bioinformatics, to determine tools regularly used by researchers and determine their aptitudes towards computer based research and learning. Using this questionnaire 50 interviews were conducted from the following establishments!

1. The National Institute of Health
2. NARC (National Agriculture Research Center)

3. Al-Shifa Hospital
4. Quaid-e-Azam University
5. The University of Arid Agriculture; and
6. Comsats Institute on Information Technology

29 of these interviews yielded significant results. The Lickert scale was applied to these observations to interpret the results. Furthermore, a variant of the Lickert scale was used to quantitatively compare the results of each interview. This produced a score representing the Bioinformatics literacy rate of each respondent.

## Introduction

Bioinformatics is a bright new field. New applications for it are being found and better ways of using them are being developed almost on daily basis. It has begun to broaden into almost all areas of research and development being carried out in Biological Sciences. Whether one speaks of Microbiology, Proteomics, Molecular Genetics, or even Zoology or Botany, Bioinformatics is changing the way research is carried out in all these fields, and in some cases even changing how the fields themselves are perceived. For example, techniques developed by computer scientists enabled researchers at the commercial Celera Genomics, the public Human Genome Project consortium, and other laboratories around the world to sequence the nearly three billion base pairs that define the 35,000 genes that constitute the human genome. Sequencing the human genome, most of which was performed from 1998 to 2000, would have been virtually impossible without the Internet; high-performance computing; and the combined efforts of mathematicians, statisticians, life scientists, physicists, and computer scientists [1]. Similarly biochemists no longer have to begin a research project by isolating and purifying massive amounts of a protein from its native organism in order to characterize a particular gene product. Rather, now scientists can amplify a section of some genome based on its similarity to other genomes, sequence that piece of DNA and, using sequence analysis tools, infer all sorts of functional, evolutionary, and, perhaps, structural insight into that stretch of DNA [2]. Databases are a necessary, integral part of this entire process. Furthermore the almost daily development of new algorithms, programs and approaches ensure that Bioinformatics itself is a changing field. Therefore there is no real consensus on the classification or its characterization. This is exemplified in the lack of a standard definition for the word [3]. Here is a sampling of definitions that we found after a simple web search. All certainly have a high degree of validity.

**Bioinformatics:** An interdisciplinary area at the intersection of biological, computer, and information sciences necessary to manage, process, and understand large amounts of data, for instance from the sequencing of the human genome, or from large databases containing information about plants and animals for use in discovering and developing new drugs [3].

**Bioinformatics:** the science of informatics as applied to biological research. Informatics is the management and analysis of data using advanced computing techniques. Bioinformatics is particularly important as an adjunct to genomics research; because of the large amount of complex data this research generates [3].

## ***The Gap between Biological and Computer Sciences:***

Students, teachers and researchers contributing to the field of bioinformatics can be divided into two major categories:

**Biological sciences Scientists** Laboratory scientists who have skills with Unix, Perl, SQL, etc., knows the limitations of available bioinformatics tools, and uses these tools and his/her scientific knowledge to help the validation process in the lab.

**Computer Programmers** who understand the biological questions being asked and have expertise in Unix, Perl, SQL, C++, Java, XML, etc. in order to build bioinformatics databases and applications [9].

The major problem with this format is the lack of communication between the two branches. This discrepancy or 'gap' prevents an appropriate merger of the two disciplines. Biological scientists do not know the required computer languages and so are unable to write custom made computer programs for themselves. Similarly computer programmers need to be informed of the requirements of the research project before they can make programs for the biological scientists to use [10].

The problem is compounded by the fact that biological scientists and computer programmers are unable to work together as a consequence of the inherent differences between the subjects. Language, methodology and conception are often very different [10]. So even when a biological scientist and a computer programmer are able to work together on a project, they are unable to achieve the idealistic synchronicity of the disciplines Bioinformatics creates.

Bioinformatics provides a unique solution to the problem by taking away both the biological scientist and the computer programmer and replacing them with one person. This 'Bioinformatician' has adequate knowledge of both disciplines and so can 'bridge the gap' completely.

## ***The situation so far:***

In developed countries around the world Bioinformatics knowledge has already been integrated in most fields related to biology. Understanding of molecular biology methods and techniques are complemented by substantial experience in searching genome databases and the analysis of DNA and protein sequence data using a variety of analytical tools and techniques. Knowledge of and the ability to use readily available web based tools and additional experience in programming, PERL/UNIX shell scripting, for example is also integrated into most courses [5].

Unfortunately this is not the case in most developing countries including Pakistan. Therefore there is a need to determine the degree and the extent of Bioinformatics literacy present in our country. Before we begin the teaching and developing our work force it is necessary to determine how Bioinformatics is used by researchers and students in their research and studies. More importantly there is a need to determine the demographic or segment of the Biological Sciences population that uses can use the current crop of Bioinformatics tools effectively [7].

This paper was designed do address this need. The term 'Bioinformatics Literacy' will be introduced. As a combination of the terms 'Computer literacy', 'Information literacy' and 'Biology' the definition we have fashioned is as follows: "Bioinformatics literacy may be defined as the ability to use information handling tools to locate, evaluate and then apply the acquired information to solve problems related to the field of Biological Sciences."

The aim this study is threefold:

1. To determine the level of Bioinformatics literacy amongst students, researchers, professionals and organizations in Pakistan.
2. To determine which Bioinformatics tools (with emphasis towards the field of Microbiology) are currently being used by various students, researchers, professionals and organizations in Pakistan.
3. To determine the reasons if any for low Bioinformatics literacy in Pakistan.

## **Materials and Methods:**

This study is a questionnaire based approach conducted on a voluntary basis. The sample set consists of research and teaching institutions in the Islamabad and Rawalpindi region. The institutions that took part in the survey are as follows:

1. The National Institute of Health
2. NARC (National Agriculture Research Center)
3. Al-Shifa Hospital
4. Quaid-e-Azam University
5. The University of Arid Agriculture; and
6. Comsats Institute on Information Technology

Firstly a list of Bioinformatics tools commonly used by the research community was prepared. Special emphasis was given to tools used in microbiology research. Using this list a questionnaire was designed. The questionnaire was designed to obtain demographics data, to assess knowledge of bioinformatics, to determine tools regularly used by researchers and determine their aptitudes towards computer based research and learning. Then the Lickert scale was applied to the results of the questionnaire.

The questionnaire itself consists of four parts. The first part measures basic parameters such as name, age gender etc. and hence reveals the demographic of the person tested. The second part determines knowledge and the ability to use Bioinformatics as determined by questions based on the list of bioinformatics tools prepared previously. The third section determines reasons for low literacy. The last section is optional and it determines the advantages of using bioinformatics.

### ***Questionnaire***

#### **Section 1**

- 1) Name
- 2) Age
- 3) Gender
- 4) Education
- 5) Current Status/ Employment

## Section 2

In this section, we asked the question “Are you aware that using Bioinformatics tools may be helpful in your work/research” from 39 different bioinformatics tools. The possible answers were:

- Have never heard of it
- Have heard of it but never used it
- Know of it and can use it
- Can use it quite well
- Know enough about it to be able to teach

The tools and resources tested in the survey are as follows:

Databases	Annotation Tools	Bacterial Genomes
Genbank	BASys	UCSC Archaeal Genome Browser
EMBL	HAMAP	xBASE
DDJB	PUMA2	Entrez Genome
BLAST	ASAP	PEDANT
UniProt	MaGe	CMR
EcoGene	IMG	Microbes online
CCDB	TIGR Annotation engine	BacMAP
Ecocyc	SABIA	
coliBASE	MAGPIE	
CCDB	GenDB	
Entrez Genome	Glimmer	
Gene Prediction	Protein Tools	Miscellaneous
GeneMark	PSORTb	Artemis
HMMER	Proteome Analyst	Bluejay
	SingalP	tRNA scan-SE
	Uniprot	EMBOSS
	Pfam	Bioperl
	tmHMM	

## Section 3

- 1) Do you have easy access to, or own a computer?
- 2) Can you effectively use a computer?
- 3) Do you have easy access to the internet?

- 4) Can you effectively use the internet (“surf the web”)?

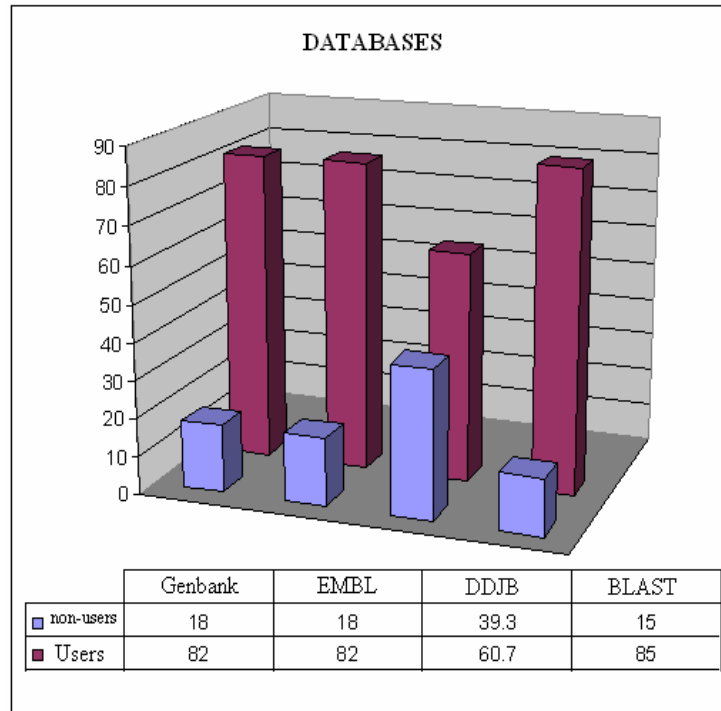
#### **Section 4**

- 1) Are there any problems or tedious approaches in your research that you think will be easier if automated / computerized?
- 2) Do you think there are any disadvantages of not using bioinformatics in your line of work?
- 3) Please mention which of the tools mentioned above are/is most useful in your work?

## Results:

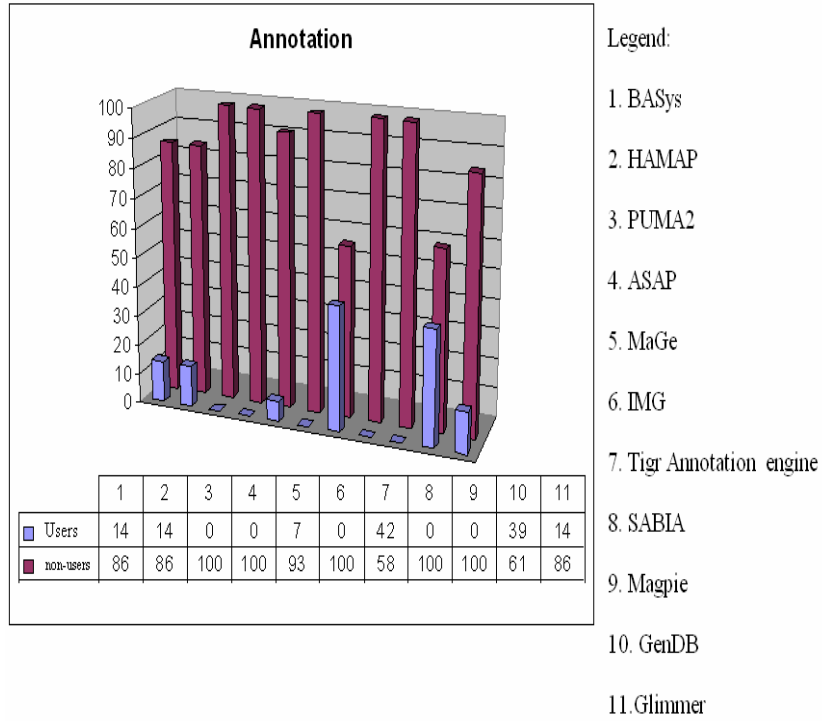
The data in this document was compiled over a 5 week period from December 22nd 2006 to February 3rd 2007. All Questionnaires were answered on a voluntary basis. Out of the 50 questionnaires handed out we were able to obtain results from 29 subjects.

A five point lickert scale was applied to the results. This divided the results into two sections, Users and Non-users. The results are as follows:



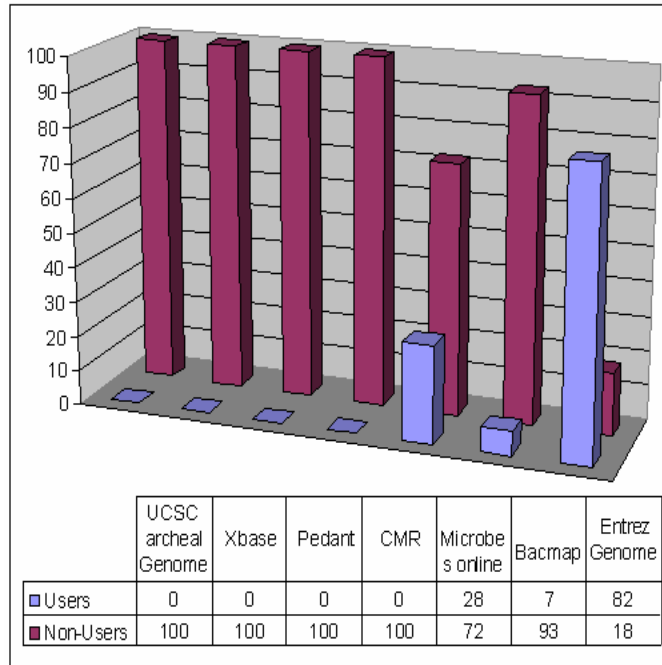
**Figure 1 Most widely used Bioinformatics Programs**

Our results show a high understanding of the most commonly used Bioinformatics tools. They are the most widely used of all the programs used. This may be because they house genomes from a wide variety of organisms.



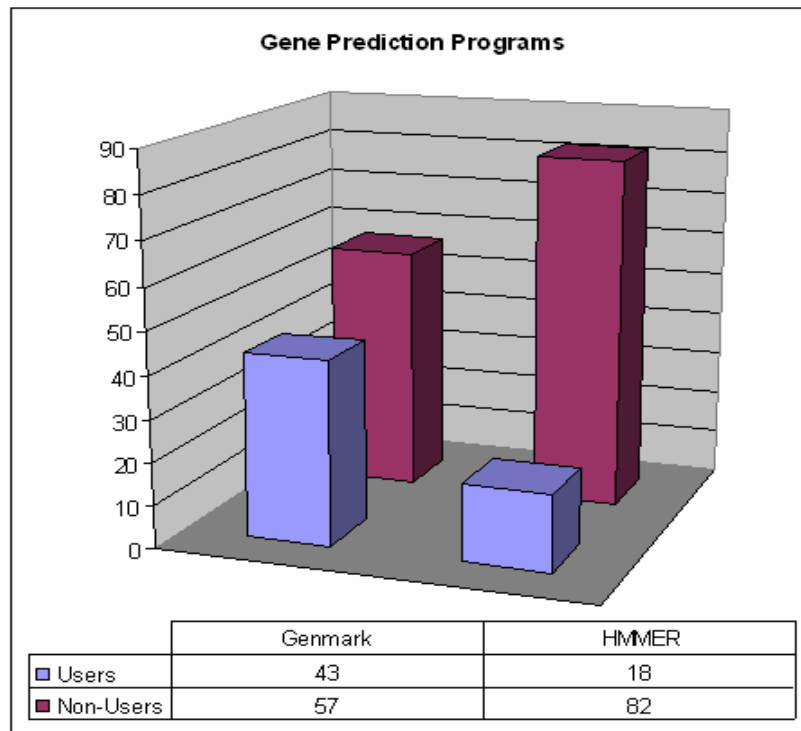
**Figure 1. 2 Chart representing knowledge of Annotation programs**

Our results show that apart from TIGR annotation engine and GenDB knowledge of annotation programs is scarce.



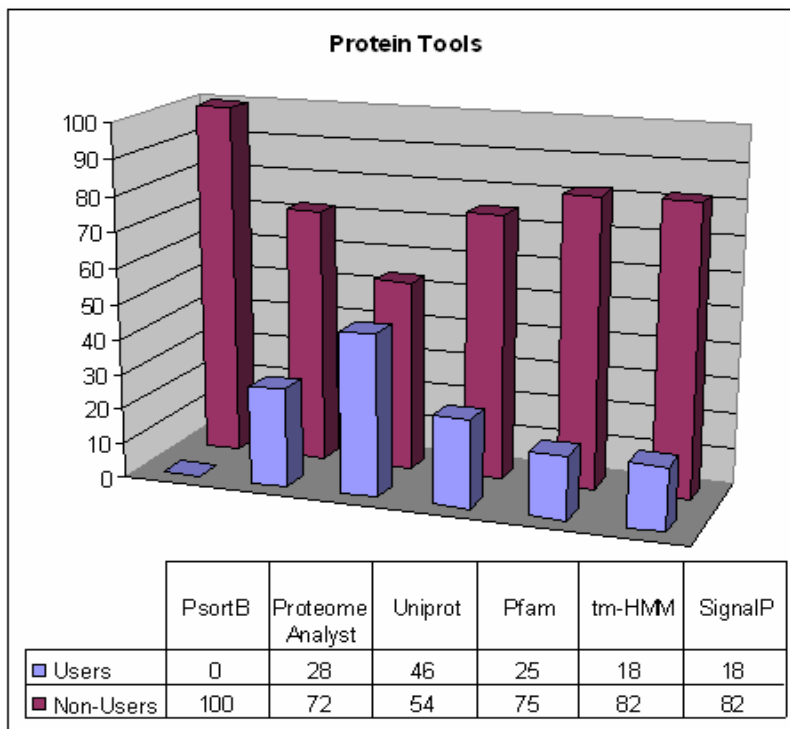
**Figure 1-3 Chart representing knowledge of bacterial genome databases**

Our results show a dearth of knowledge regarding bacterial genome databases. This is surprising since a majority of our interviewees were of microbiological background. This suggests that the microbiologists are using Entrez Genome, a database used to provide genome data on a variety of organisms more than biological databases.



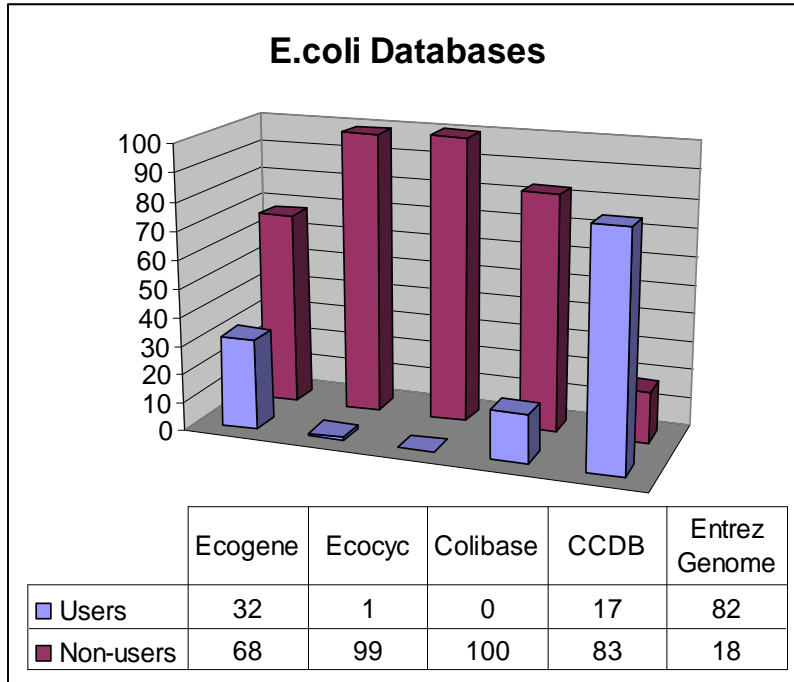
**Figure 1-4 Chart representing use of Gene Prediction Programs**

Our research shows a low tendency towards using gene prediction tools. This may be because this type of research wasn't being carried out at the time of the survey.



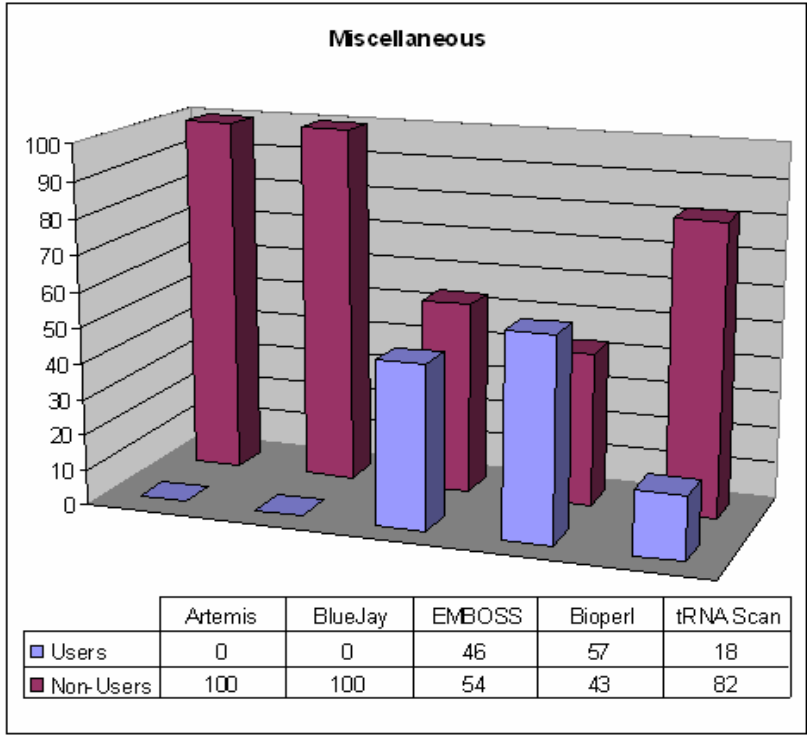
**Figure 1-5 Chart Representing use of Protein Tools**

Apart from Uniprot, a database used to gather data on proteomes, no biological tools are used. This is not surprising since a majority of the interviewees were not working on protein based projects at the time this survey was conducted.



**Figure 1-6 Chart representing use of bacterial genome databaes**

A majority of the interviewees do not use bacterial genome databases. Therefore it is not surprising to see that E. coli databases are not used as well.



**Figure 1.7 Miscellaneous Usage of various Bioinformatics utilities**

Our results suggest that a large number of the interviewees can use bioperl. We were unable to determine however, whether they are able to program and so create custom made programs.

**Bioinformatics Literacy Data:**

A variant of the lickert scale was used to quantitatively determine the bioinformatics literacy level of each interviewee. The results are as follows:

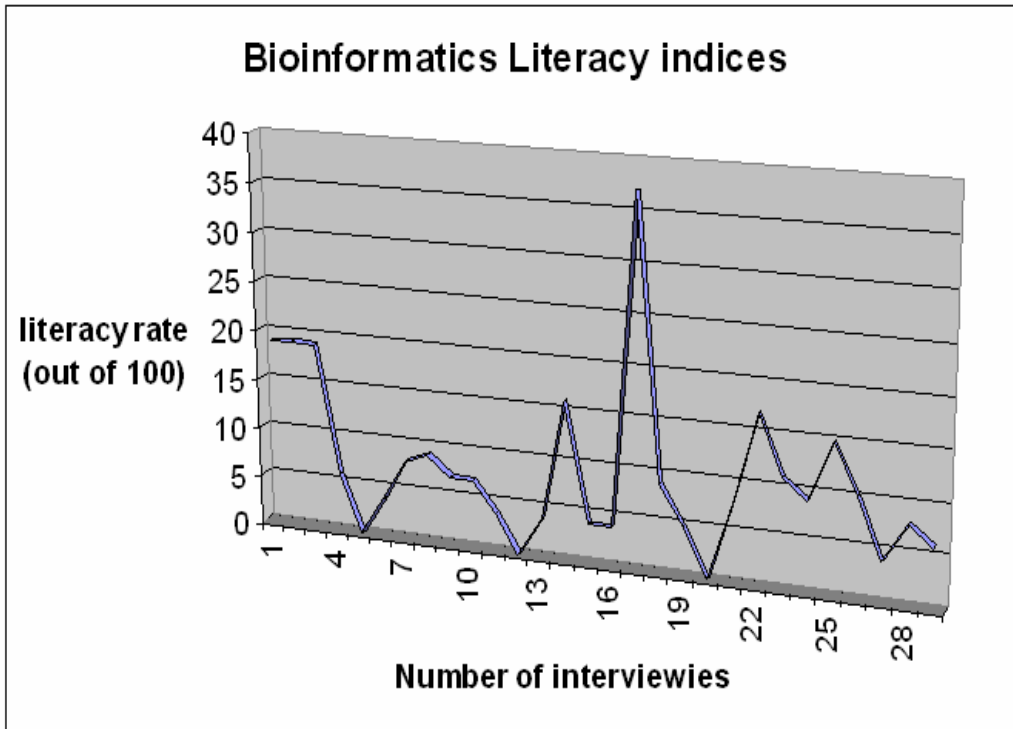


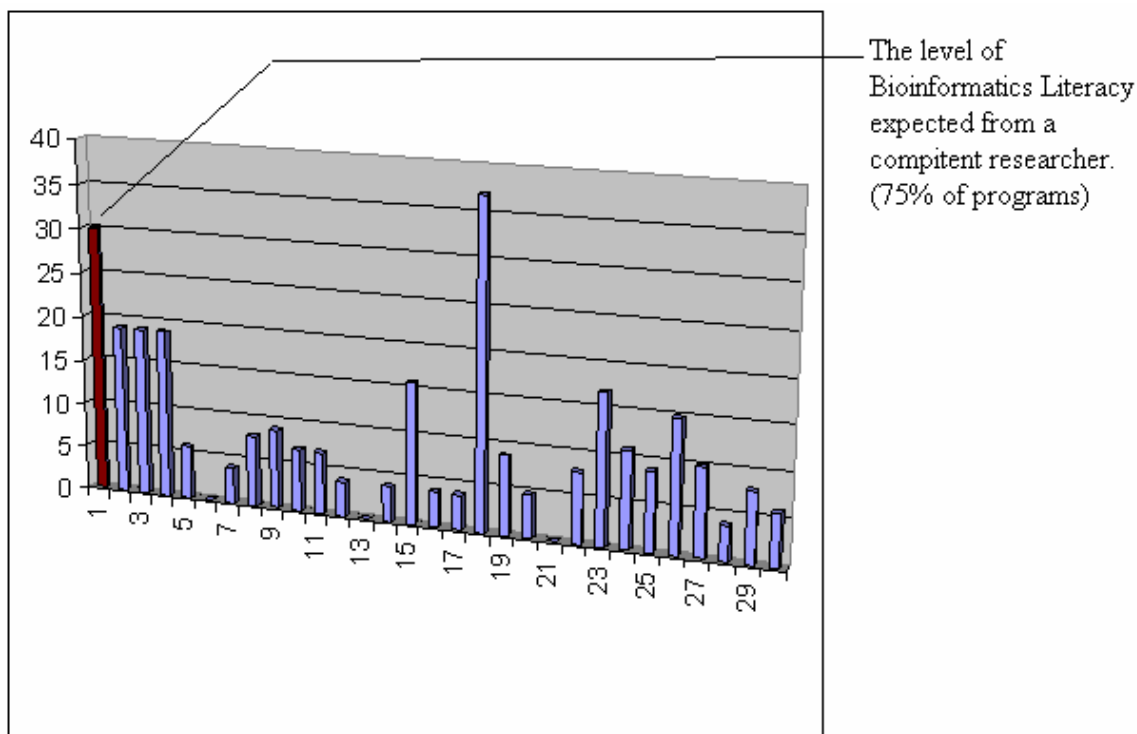
Figure 1.8

## Conclusions:

Bioinformatics is an emerging field. In Pakistan knowledge and understanding of the subject is scarce. This is exemplified by the fact that the first bioinformatics graduates in Pakistan only passed out within the last year. Over the course of this project we interviewed researchers and students from a wide variety of backgrounds in technical expertise. Unfortunately the results showed a very low degree of bioinformatics literacy.

### ***Bioinformatics Literacy rates:***

To calculate the degree of Bioinformatics skills the results from each questionnaire were quantized using a custom built variant of the Lickert scale. The results show a low understanding of web based bioinformatics tools. This is exemplified by the following graph.



### ***The tools used by students and professionals:***

This table offers a breakdown of the tools used.

<b>Name of Databases</b>	<b>No. of people using them</b>
1) GenBank	<b>15</b>
2) EMBL	<b>7</b>
3) Entrez Genome	<b>6</b>

4) PSORTb	2
5) Proteome analyst	3
6) BLAST	17
7) Uniprot	2
8) Pfam	3
9) EMBOSS	3
10) Bioperl	3
11) Justbio	1
12) GenScan	1
13) ORF finder	2
14) CLUSTAL W	2
15) Swiss PDB	1
16) Modeller	1
17) Protein dock	1
18) Gene mark	4
19) tRNA Scan	2
20) Restriction analysis program	1
21) Gene prediction program	1
22) RNA secondary structure analysis	1
23) Phlylogenetic phylip	1
24) Microbes online	2
25) xBASE	2

26) DDBJ	2
27) ECOgene	1
28) Artemis	1
29) Bluejay	1
30) HMMR	1
31) Signal IP	1
32) SABIA	1
33) TMHMM	1
34) DNA seq. multiple alignment	1
35) TIGR	2
36) Expasy	1
37) UCSC	1
38) MEGA 3	2
39) Cn3d	2
40) HMMER	1
41) Primer3	1

***Reasons for low literacy:***

The results show that a majority of the people interviewed have access to a computer and the internet. Furthermore, the results show that they are capable of using both effectively. This suggests that the lack of bioinformatics knowledge is not because of a dearth of resources and infrastructure.

We hypothesize two reasons for the low levels of literacy. Firstly, a lack of awareness, the average researcher does not use bioinformatics for his work because he does not know of them. The second reason is that they may be using those tools not tested for in our survey.

## Appendix A: URLs of Bioinformatics Tools

- 1) EcoCyc <http://ecocyc.org/>
- 2) PUMA2 <http://compbio.mes.anl.gov/puma2/>
- 3) Colibase <http://colibase.bham.ac.uk/>
- 4) CCDB <http://redpoll.pharmacy.ualberta.ca/CCDB/>
- 5) Enterz genome <http://eutils.ncbi.nlm.nih.gov/enterz/query.fcgi?db=genome/>
- 6) xbase <http://xbase.bham.ac.uk/>
- 7) Pedant <http://pedant.gsf.de/>
- 8) CMR <http://cmr.tigr.org/tiger-scripts/CMR/CmrHomePage.cgi>
- 9) microbesonline <http://www.microbesonline.org/>
- 10) BacMap <http://wishart.bioiology.ualberta.cd/BacMap/>
- 11) Psort <http://www.psort.org/psortb/>
- 12) Pfam <http://www.sanger.ac.uk/Software/Pfam/>
- 13) tmHMM <http://www.cbs.dtu.dk/services/TMHMM>
- 14) SignalP <http://www.cbs.dtu.dk/services/SignalP>
- 15) BASys <http://wishart.bioiology.ualberta.ca/basys/>
- 16) HAMAP <http://ca.expasy.org/sport/hamap/>
- 17) MaGe <http://www.genoscope.cnsfr/age/mage/>
- 18) IMG <http://img.jgi.doe.gov/>
- 19) Magpie <http://magpie.ucalgary.ca/>
- 20) GenDB <http://www.cebitec.uni-bielefeld.de/groups/brf/software/gend/info>
- 21) BLAST <http://www.ncbi.nlm.nih.gov/blast>
- 22) Bluejay <http://wwwbluejay.ucalgary.ca/>
- 23) EMBOSS <http://emboss.sourceforge.net/>
- 24) tRNAScan-SE <http://selab.wustt.edu/>
- 25) UCSC Archaeal Genome Browser <http://archaea.ucsc.edu/>
- 26) GLIMMER <http://www.tigr.org/salzberg/glimmer.html/>
- 27) Artemis <http://www.sanger.ac.uk/software/Artemis/>
- 28) Ecogene <http://ecogene.org/>
- 29) GeneMark <http://exon.gatech.wustt.edu/>
- 30) Bioperl <http://www.bioperl.org/>
- 31) Proteome Analyst <http://www.cs.ualberta.ca/bioinfo/PA>
- 32) Uniprot <http://www.pir.uniprot.org/>
- 33) ASAP <http://asap.ahabs.wise.edu/asap/ASAP1.htm/>

## References:

- [1] Human Genome Project [www.khwarzimic.org/takveen/genome-dawn.pdf](http://www.khwarzimic.org/takveen/genome-dawn.pdf)
- [2] A Bioinformatics Survey, Steve Thompson, Florida State University School of Computational Science and Information Technology (CSIT) (ppt)
- [3] What is bioinformatics? (book)
- [4] Applied Bioinformatics Computing: An Introduction, By Bryan Bergeron.
- [5] Computing Concepts for Bioinformatics, Nirav Merchant and Susan Miller (ppt)
- [6] Bioinformatics I, An introduction And where to find biological information (ppt)
- [7] Developing Bioinformatics Computer Skills, Cynthia Gibas & Per Jambeck  
April 2001 (book)
- [8] Bioinformatics what and why, modified from information supplied by Dr. Bruno Gaeta (ppt)
- [9] Opportunities for the Pakistani IT industry within Bioinformatics, Gallup report
- [10] Delivering Bioinformatics Training: Bridging the Gaps Between Computer Science and Biomedicine, Christopher Dubay Ph.D., James M. Brundege Ph.D., William Hersh M.D. Kent Spackman M.D., Ph.D.
- [11] Developing a Core Bioinformatics Syllabus, Dr. Natalio Krasnogor, School of Computer Sciences & Information Technology (ppt)