Patron-in-Chief : Prof. Sangram Mudali
Director, National Institute of Science & Technology
Palur Hills, Berhampur-761 008, Orissa, India

Editor-in-Chief : Prof. Suash Deb
Dept. of Computer Science & Engineering,
National Institute of Science & Technology
Palur Hills, Berhampur-761 008, Orissa, India
E-mail : suashdeb@yahoo.com

Editorial Board :
Lory Satpathy
Rajanikanta Das
Yashodhara Das
Ipshita Sahu
Bismaya K. Purohit
Aditya Ranjan Padhi
## CONTENTS

1. Message from the Director, NIST 01
2. Message from the Editor-in-Chief 03
3. Message from Dr. Debesh Das, Hon’ble Minister, I.T., Govt. of West Bengal 05
4. Message from Shri P. Ray, Addl. Chief Secretary, Govt. of West Bengal 07
5. Message from Dr. T. Ramasamy, Secretary, Dept. of Science & Technology, Govt. of India 09
6. Message from Mr. Prasun Mukherjee, Commissioner of Police, Kolkata 11
7. Bioinformatics — Practice and Perspectives, Part-IV 13
8. Databases and Browsing Tools for Gene Prediction 16
10. Company Profile—Jubilant Biosys Ltd., Bangalore 24
11. Info-NIST 25
12. Alumnis’ Delight at Infosys 26
13. Student’s Voice 27
14. Readership Survey 28
MESSAGE FROM THE DIRECTOR, NIST

Dated: August 5, 2007

Dear Friends,

Thanks very much for your overwhelming feedback vis-a-vis my concern, as expressed in the last issue (May’07) of bioinformatika regarding the growing reluctance of our students to embrace R & D. It is often being heard that the placement record of NIST is indeed, enviable. But at the same time acting as a placement agency in the guise of an engineering & management institute runs contrary to the NIST philosophy. Here placement had never been an end in itself. Rather the emphasis is on making our students competent enough so as to produce really deserving candidates for getting jobs. By job I don’t intend to emphasize merely on the brand name of organizations or that of very healthy pay packets. Rather my focus is on how challenging & how innovative the assignments are. Viewing from that angle, I found our students’ antipathy for R&D intriguing & disappointing.

One of the newsworthy information of recent times had been the decoding of the genome of the “father of DNA”, Dr. James D. Watson. The deciphering of the genome of Dr. Watson, who jointly discovered the structure of DNA in 1953, is believed to mark the gateway to an era of personalized genomic medicine.

Making somebody’s genome publicly available is fraught with considerable risks. It might reveal unpleasant information about that person as well as that of their kith & kin. But Dr. Watson, a Nobel laureate, dedicated the entire genome to the researchers with the single exception of his apolipoprotein E gene, the status of which he refused to know as it predisposes a person toward Alzheimers disease.

This issue of bioinformatika is the last one before the completion of its one year journey. If the opinion of experts are of any indication, the decision to launch the same has been justified to the full extent through the superlative effort of the team of bioinformatika. Amongst others, the excitement and pride amongst our (2007 pass out) students, currently undergoing training at Infosys Technologies Ltd, Mysore DC, (on discovering Bioinformatika occupying space at the Infosys library) is truly noteworthy. While thanking them for sharing their excitement with us. I would like to see the team of bioinformatika making pledge towards carrying it to more heights in the coming days and thereby bringing more exuberance for the entire NIST fraternity for it.

With Best Wishes,

(Sangram Mudali)

(Prof. Sangram Mudali)
MESSAGE FROM THE EDITOR-IN-CHIEF

Dated : July 31, 2007

Hello Colleagues,

The last issue (May’07) of Bioinformatika saw my article (entitled : “2007 World Cup Cricket & Bioinformatika : Some Lessons”) raising the issue amongst others, about how can bioinformatics be made more popular amongst the students, particularly those pursuing B.Tech./BE/ M.Tech in IT/Computer Science. I am happy to report that the same had been able to attract sizeable responses. A subset of those responses attempted to draw some conclusions vis-a-vis the importance of bioinformatics by taking into account the statistics of student enrollment. While thanking all the respondents kindly permit me to differ with this particular school-of-thought & share my personal view with the readers of bioinformatika on the same.

The importance of a subject/course, I firmly believe, can never be proportional to student enrollment i.e. how many students are enrolling in that particular course. Any attempt to do so is indeed, meaningless since it endeavors to trying to rank subjects through popular votes (i.e. the student strength). It runs completely against the premises of a scientific temper.

While never making any attempt to ignore or undermine sentiments, liking/disliking of the student community, it can be safely argued that students will always tend to opt for subjects/courses for which the job market is very hot & also for which the effort required to pass/securing high marks as well as getting established in life & climbing corporate ladders are relatively less painful & less cumbersome. This tendency of striking the iron when it is hot might undoubtedly throw some light on the popularity of that (hot iron) course/subject at that particular point of time but can never undermine the importance of the others. Gimmicks cannot advance interest in any subject. Albeit not mutually exclusive, importance & popularity of a course/subject, unlike in sports or movies, do not always point towards the same direction. Popularity contest based on the feedback of the students, once introduced amongst different subjects of science & engineering, will be simply silly & ridiculous.

Let us now turn our attention to this issue of bioinformatika. Every issue of the magazine is witnessing an increased enthusiasm for submission of article(s) and the Aug’07 one was no exception. It contains very interesting articles from diverse regions of our country. Thanks to all of the contributors for their efforts. In
addition, this issue is privileged to receive Good Wish messages from very distinguished personalities like i) Hon’ble Minister of Information Technology, Govt. of West Bengal, ii) Additional Chief Secretary - Govt. of West Bengal, iii) Secretary - Dept. of Science & Technology (Govt. of India) & iv) Commissioner of Police - Kolkata. The inspiring messages from such personalities has no doubt, enhanced the enthusiasm of the entire team of the magazine to a very great extent. On behalf of the editorial team, I hereby convey my thanks & gratitude to all of them for their patronage & seek their continued guidance for the magazine. Also the messages from some of our students (2007 passout) currently placed at Infosys are indeed, encouraging, especially for the student members of the editorial board as well the members of The NIST Bioinformatics Club. However, numerous other feedback/messages as received from various places and which we found extremely interesting as well as educative could not be printed in this issue due to space constraints. Trust we won't be misunderstood.

Starting with this issue, we are undertaking a formal readership survey. Trust the same will be more helpful towards feeling the pulse of our reader in a more comprehensible way. Kindly spare some moments of your valuable time for filling up the readership survey form (attached at the end of the magazine) and send it to me through the mode (electronic or otherwise) most convenient for you.

The journey of bioinformatika is nearing its completion of 1 year. The next issue (Nov’07) will mark the 1-st anniversary of this mouthpiece of The NIST Bioinformatics Club. Please feel free to establish contact with me & communicate with your views about how the 1-st anniversary issue can be made something special for our readers. Of course, your inputs help us towards making every issue of bioinformatika special. But let us focus on the Nov’07 issue of the same for now.

Awaiting your more enthusiastic response & participation for the 1st anniversary issue of bioinformatika,

With Warm Regards,

Yours sincerely

(Prof. Suash Deb)
MESSAGE FROM DR. DEBESH Das

May 21, 2007

To
Prof. Suash Deb
Editor-in-Chief, Bioinformatika,
Dept. of Computer Science & Engineering,
National Institute of Science & Technology,
Pular Hills,
Beharampur - 761 008, Orissa.

Sub : Bioinformatika

Dear Suash,

I am indeed, delighted to go through the magazine captioned "Bioinformatika" which, I believe, serves as platform for exchanging views among bioinformaticians in India and also spreading upto date information on bioinformatics to the common people and thereby educating them.

I congratulate all concerned with this magazine for their commendable job and wish all the success to their endeavour.

Yours sincerely

(Dr. Debesh Das)
MESSAGE FROM SHRI P. RAY

No. 552-HSPA

May 24, 2007

Dear Suash,

It has been a great pleasure for me personally as also professionally to glance through the three issues of Bioinformatika that you sent me. I was deeply impressed with the production of the magazine and its contents. I found that you are indeed, covering path breaking ground in India.

Bioinformatics has many application areas and some of them sound quite esoteric to the uninitiated. However, there are areas which are proving extremely important for detection of crimes. Pattern recognition has been used for quite sometime as a tool of forensic science. A new dimension that has opened up recently is the DNA analysis facility. As yet, its application has been restricted, but in the days to come, this could prove to be a major tool in detection, arrest and conviction of crime perpetuates. I believe that Bioinformatics can act as a really effective tool for fighting crimes.

Bioinformatika, which you have been handling quite ably can certainly shed more light of this and create awareness regarding the use of Bioinformatics not only amongst the common people but also amongst the police officers who need greater exposure to such ideas.

I congratulate The NIST Bioinformatics Club and the entire team of Bioinformatika for the tremendous efforts that have been put in and wish this venture a grand success.

With Best Wishes,

Yours sincerely

(P. Ray)

To,

Prof. Suash Deb
Editor-in-Chief, Bioinformatika
National Institute of Science & Technology
Palur Hills, Berhampore 761008, Orissa

Shri P. Ray
IAS

ADDITIONAL CHIEF SECRETARY
HOME DEPARTMENT
GOVERNMENT OF WEST BENGAL
Writers’ Buildings, Kolkata-700 001
Tel. : 2214-5656, Fax : 2214-3001
MESSAGE FROM DR. T. RAMASAMI

MESSAGE

The publications of The NIST Bioinformatics Club bridge an important area of need. On behalf of Department of Science and Technology, I extend hearty congratulations to you and your team. It is an important and valuable initiative.

Bioinformatika is really a medium that Bioinformaticians of India would have been looking for. It offered a forum through which they could bridge themselves and together reach common people. The publication is both timely and valuable.

To,
Prof. Suash Deb
Editor-in-Chief, Bioinformatika
Department of Computer Science & Engineering
National Institute of Science & Technology
Palur Hills, Berhampore 761008, Orissa

Tel. : 0091-11-26510068 / 26511439, Fax : 0091-11-26863847 / 26862418
MESSAGE FROM MR. PRASUN MUKHERJEE

Prasun Mukherjee I.P.S.
Commissioner of Police
Kolkata

18, Lalbazar Street,
Kolkata 700 001,
2214 5060 (O)
2214 5424 (F)

To,

Prof. Suash Deb
Editor-in-Chief, Bioinformatika
Dept. of Computer Science & Engg.
National Inst. of Science & Technology
Palur Hills, Berhampore 761008
ORISSA

June 7, 2007

Sub : Bioinformatika

Dear Suash,

It was really a great pleasure to see your effort in the field of bioinformatics and the publication of the quarterly magazine -Bioinformatika. I was impressed to notice the quality of the magazine and its contents.

Bioinformatics has got myriads of application areas. One such domains had been the DNA analysis facility for facilitating the detection of crimes. In the technology era, the perpetrators of crime are also increasingly going high tech for carrying out various subversive activities. This is making the detection and arrest of perpetrators more & more difficult.

Bioinformatics, I strongly believe can act as an effective tool for fighting this.

Bioinformatika, under your stewardship, can definitely throw more light on this and other aspects & create an awareness amongst the common people.

I congratulate The NIST Bioinformatics Club & the entire team of Bioinformatika for the tremendous effort & wish it a grand success.

With best wishes,

(P. Mukherjee)

7/6/07
Some major areas of research in bioinformatics:

Drug Design: The information present in DNA is expressed via RNA molecules into proteins which are responsible for carrying out various activities. This information flow is called the central dogma of molecular biology (Fig. 7). Potential drugs can bind to DNA, RNA or proteins to suppress or enhance the action at any stage in the pathway.

As structures of more and more protein targets become available through crystallography, NMR and bioinformatics methods, there is an increasing demand for computational tools that can identify and analyze active sites and suggest potential drug molecules that can bind to these sites specifically (Fig. 8).

Table 4: Leading causes of death

<table>
<thead>
<tr>
<th>Leading causes of death in millions per year due to infectious diseases (for the year 2002)</th>
<th>Percentage (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lower respiratory infections</td>
<td>3.9</td>
</tr>
<tr>
<td>HIV/AIDS</td>
<td>2.8</td>
</tr>
<tr>
<td>Diarrheal diseases</td>
<td>1.8</td>
</tr>
<tr>
<td>Tuberculosis</td>
<td>1.6</td>
</tr>
<tr>
<td>Malaria</td>
<td>1.2</td>
</tr>
<tr>
<td>Malaria</td>
<td>0.8</td>
</tr>
<tr>
<td>Neonatal Causes</td>
<td>NA</td>
</tr>
<tr>
<td>Other (including noncommunicable diseases)</td>
<td>NA</td>
</tr>
</tbody>
</table>

Time and cost required for designing a new drug are immense and at an unacceptable level. According to some estimates it costs, on an average, about $880 million and 14 years of research to develop a new drug before it is introduced in the market (Table 5).

Table 5: Cost and time involved in drug discovery

<table>
<thead>
<tr>
<th>Stage</th>
<th>Duration</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Target Discovery</td>
<td>2.5 yrs</td>
<td>4%</td>
</tr>
<tr>
<td>Lead Generation &amp; Lead Optimization</td>
<td>3.0 yrs</td>
<td>15%</td>
</tr>
<tr>
<td>Preclinical Development</td>
<td>1.0 yrs</td>
<td>10%</td>
</tr>
<tr>
<td>Phase I, II &amp; III Clinical Trials</td>
<td>6.0 yrs</td>
<td>68%</td>
</tr>
<tr>
<td>FDA Review &amp; Approval</td>
<td>1.5 yrs</td>
<td>3%</td>
</tr>
<tr>
<td>Drug to the Market</td>
<td>14 yrs</td>
<td>$880 million</td>
</tr>
</tbody>
</table>

(Source: PAREXEL [21])

Intervention of computers at some plausible steps is imperative to bring down the cost and time involved in the drug discovery process (Table 6). Making a drug is more like designing a key for a lock to jam or open the lock, except that both the lock and the key are dynamic and made of atoms and are susceptible to environmental effects such as solvent, salt and other small or biomolecules.

Table 6: High End Computing Needs for In Silico Drug Design:

In silico methods can help in identifying drug targets via bioinformatics tools. They can also be used to analyze the target structures for possible binding / active sites, generate candidate molecules, check for their drug-likeness, dock these molecules with the target, rank them according to their binding affinities, further optimize the molecules to improve binding characteristics etc. Some popular softwares for drug design are listed in Table 7.

Table 7: Comprehensive softwares for drug design

<table>
<thead>
<tr>
<th>S.No</th>
<th>Name of the software</th>
<th>Description</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>InsightI, Discovery studio Cerius2 ADME/Tox Package</td>
<td>A suite for molecular modeling and de novo drug design Provides computational models for the prediction of ADME properties derived from chemical structures</td>
<td><a href="http://www.accelrys.com/products/insight/">http://www.accelrys.com/products/insight/</a></td>
</tr>
<tr>
<td>2</td>
<td>Sybyl</td>
<td>Computational informatics software for drug discovery</td>
<td><a href="http://www.sybyl.com/">http://www.sybyl.com/</a></td>
</tr>
<tr>
<td>3</td>
<td>Bio-suitie</td>
<td>The package can be used for Genomics, Protein modeling &amp; Structural analysis, Simulation and Drug Design</td>
<td><a href="http://www.ncbi.nlm.nih.gov/standards/genesuite/">http://www.ncbi.nlm.nih.gov/standards/genesuite/</a></td>
</tr>
<tr>
<td>4</td>
<td>Molecular Operating Environm ent (MOE)</td>
<td>Cheminformatics, Protein Modeling, Structure-Based Design, High Throughput Discovery and Molecular Modeling and Simulations</td>
<td><a href="http://www.chemcomp.com/">http://www.chemcomp.com/</a></td>
</tr>
<tr>
<td>5</td>
<td>MDL-QSAR</td>
<td>A suite of tools and applications for decision support, visualization, analysis, and ADME-toxicity assessment</td>
<td><a href="http://www.mdl.com/products/predictive">http://www.mdl.com/products/predictive</a> QSAR/index.jsp</td>
</tr>
<tr>
<td>7</td>
<td>Autodock</td>
<td>Protein +ligand docking</td>
<td><a href="http://www.scripps.edu/molbio/sonido/autodock/">http://www.scripps.edu/molbio/sonido/autodock/</a></td>
</tr>
<tr>
<td>8</td>
<td>Ligand</td>
<td>Program for automatically plotting protein ligand interactions</td>
<td><a href="http://www.chemlab.ucsd.edu/ligand/ligand.html">http://www.chemlab.ucsd.edu/ligand/ligand.html</a></td>
</tr>
</tbody>
</table>
Pursuing the dream that once the gene target is identified and validated, drug discovery protocols could be automated using bioinformatics and computational biology tools, we have developed a computational protocol for active site directed lead molecule design. The suite of programs christened “Sanjeevini” (http://www.scfbio-iitd.res.in/research/drugdesign.htm)[22] has the potential to evaluate and/or generate lead-like molecules for any biological target [23, 24]. The Sanjeevini protocols when tested on Cyclooxygenase-2 and Estrogen receptor as targets could successfully distinguish drugs from non-drugs (Fig. 9). Validation on other targets is in progress.

Bioinformatics applied in the form of pharmacogenomics involves developing personalized medicine for individuals based on their genetic profile. Databases of genetic profiles of patients with ailments like diabetes, cancer etc. play an important role in individual health care. The aim is to study a patient’s individual genetic profile and compare it with a collection of reference profiles which may help in improving the diagnosis and prevention of the disease.

Bioinformatics endeavors in India

Owing to the well acknowledged IT skills and a spate of upcoming software, biotech and pharma industries and active support from Government organizations, the field of Bioinformatics in India appears promising. However, the projections of the growth potential in India in a global scenario clearly indicate (Fig. 11) that a lot more could be done.
Introduction
The human genome sequence represents the first bounded biological dataset concerning our species. Having access to it is a landmark because of the limits it sets on the problem of understanding biology as a whole. It has allowed us to assemble something equivalent to an “edge” of a multidimensional jigsaw puzzle. It is the first step toward other complete datasets: the complete set of genes, proteins, molecular interactions in the cell, etc. Genome sequences provide a framework around which all this biological knowledge can potentially be organized, so each layer of data will lead to a greater understanding of layers of organization of biological systems above it.

The availability of several closely related genome sequences (e.g., mouse, rat) and evolutionary similarities between individual proteins brings the possibility of building lists of molecular features common to all vertebrate species and those that are unique to our own. Other more distant non vertebrate genomes cannot be usefully compared to human at the chromosomal level, but orthologous genes and proteins can be identified. Such analysis will increase our understanding on what makes species different from each other and also how biology accommodates change through evolution within populations of individuals.

All biological data which includes genome sequences, protein structures, RNA expression patterns, and cellular localization images, has created a huge need for databases to store information, provide access, and add value. For a current snapshot of the huge range of biological databases a good source is the annual special database issue of Nucleic Acids Research, published each January. It lists 339 databases in its opening review article in 2002 (2).

Long before the first complete genome sequences of free living cells were determined, groups from around the world had been tackling the issues of (a) building repositories for raw data, (b) adding annotation to this raw data, and (c) providing higher level structure and organization. Examples of repositories are the public DNA sequence databases of EMBL, GenBank (6), and DDBJ (10) as well as the public protein structure database PDB (11). Examples of annotation databases include Fly base and SwissProt. Examples of organizational databases are Pfam, which groups protein sequence domains into families, thereby showing evolutionary relationships between paralogous proteins within an organism and orthologous proteins between organisms.

A major current challenge for all these database projects is to increase their integration by means that may include propagating information upward from the complete genomes. One of the problems that urgently need to be addressed in this integration is the maintenance of evidence trails linking derived annotation with the source of its evidence. For example, a protein of unknown function is labeled as a kinase because of a weak sequence homology to another protein that is known to be a kinase. Later it is discovered that the weak homology between the sequences was false and was due to a frame shift error in one of the protein sequences. Because most
databases do not track the relationship between annotation and the evidence that supported it, the ‘kinase’ label is likely to persist even when the justification for it has vanished. A perfect integration would allow biology to be viewed in vertical slices (from genome sequence to organism) and horizontal slices (from viruses to humans). A vertical slice might be the biology around a gene while a horizontal slice might be the biology around a metabolite.

From every standpoint, we are a very long way from attaining such a goal: in terms of availability of experimental data, in terms of being able to predict the behavior of biological systems, and in terms of organizing existing data. However the initial steps are very clear. This review will concentrates on the issues of surveying the systems that are being constructed to automatically annotate the human genome for genes and the computer systems technology that are being assembled to store, manipulate, and provide access to these very large blocks of data.

**Automatic Gene Annotation Analysis Pipelines**

The human and mouse genomes are currently much larger than any other genome sequenced, and the sheer size of the data has posed many problems to those trying to annotate it. These problems have been partially solved by increasing the number of computers available for the analysis, but still large portion of the analysis has to be automatic to get acceptable annotation turnaround, and this involves engineering an analysis pipeline system to process the vast amount of sequences. The early stages of the different pipelines are very similar; before any gene annotation can proceed many different programs need to be run to provide input data for the gene-building system. The genomic sequence needs to be masked for repeats [RepeatMasker (9)] before doing most other analyses. After the repeat masking the next stage is to run both abinitio gene-prediction programs (Genscan Genie) and database searches against the public databases. The point at which this “raw analysis” is done differentiates various genome pipelines. Each site uses different programs and sources of data to produce their gene annotations, and this part of the analysis is the most dynamic and subject to change.

**Gene-Prediction Methods**

Gene-prediction methods have traditionally been described as falling into two distinct approaches. The first approach has been to use the statistical information contained within the genomic sequence to predict gene structures. There are many programs that perform this ab initio prediction, including Genscan, Fgenesh, HMMGene (7), Grail (12), and Genie. The algorithms behind these programs are Hidden Markov Models (HMMs), neural networks, and linear discriminant functions; although the majority now uses HMMs. But all abinitio methods have a common weakness for large genomes. The amount of coding sequence compared to noncoding sequence is very small (approximately 2% of the human genome is estimated to be coding), which leads to large introns and large intergenic regions with relatively small exons. For instance, the mean size of an exon in the human genome is 120 base pairs with the median intron size being 2000 base pairs. The large noncoding regions lead the ab initio programs astray and lead to extra genes being predicted in intergenic regions.

The second approach to gene prediction has used the wealth of sequence data from genes that is already stored in the public sequence databases which are generally similarity-based approaches founded around taking a protein or a cDNA and aligning it to the genomic sequence. There are many programs that will do a local or global alignment of two sequences, but
Bioinformatika

for gene annotation these programs are not good enough as they do not have any knowledge of gene structure. For accurate alignment, a program should be able to model splice sites, exons, introns, and in the case of protein alignment programs, open reading frames. For aligning cDNAs to genomic sequences, programs such as ESTGENOME (8), BLAT (J. Kent, manuscript in preparation), and Exonerate (G. Slater & E. Birney, manuscript in preparation) are commonly used. Gene wise and Procrustes are used for aligning proteins.

Sources of Genome Annotation

There are currently three main sites providing free annotation of the human genome. These are the Ensembl site (http://www.ensembl.org), NCBI (http://www.ncbi.nlm.nih.gov), and the genome browser at University of California, Santa Cruz (UCSC) (http://genome.cse.ucsc.edu).

ENSEMBL: The Ensembl gene-prediction system uses both abinitio and similarity-based methods but increases the specificity of the abinitio predictions by only using the ones confirmed by similarity to protein, cDNA, or expressed sequence tag (EST) sequences. Ensembl predicts genes in three steps: by placing known human genes onto the genome, then placing highly similar genes (maybe from mouse or rat) onto the genome, and finally predicting novel genes from abinitio predictions supported by sequence similarity. The final alignment of each protein is done using the Gene wise system, which uses an HMM to model exons, introns, and their splice sites. Genewise is a very slow program and the genome is very large, each protein to be aligned is placed on the genome using two methods. For the human proteins a very fast exact-matching algorithm, pmatch (R. Durbin, unpublished) or Exonerate, is used to place each protein at its approximate place in the genome. To refine the location of the protein, BLASTX is then used. This should locate the positions of the exons to within a few bases and also picks up small exons for which pmatch are not sensitive enough. Only then is Genewise used to align the protein. The Genewise/Exonerate method for predicting genes accounts for 70 percent of the Ensembl genes.

NCBI: The gene predictions on the NCBI website are based around the Genome scan program (13), which is an extension of the Genscan program that uses a probabilistic model of exon-intron structure and compositional features of human genes. Genomescan combines abinitio gene predictions with similarities to protein sequences in order to predict gene structures that have at least one exon with supporting evidence from an existing protein sequence. The final gene structure predicted by Genomescan is based on the presence or absence of regions of sequence that are similar to known proteins. These regions of similarity can be generated using any method the user prefers. BLASTX program is a popular method and is capable of generating the similarity regions. Each region is evaluated and assigned a probability which the Genomescan compares with HMM-state probabilities, thus leading to predicted gene structures that are more likely than not to include the similarity regions.

Genome Browser

The UCSC human-genome site provides multiple sets of gene predictions, some of which are produced at UCSC and some of which are provided by external groups. The group at UCSC produces a set of genes originating from the human section RefSeq. The alignment of RefSeq cDNAs is done with BLAT which is designed to find high-similarity (>95%) matches of lengths of 40 bases or more. Aligning RefSeq genes to a genome will only provide a subset of all human genes present in the genome.

Another set of gene predictions on the UCSC browser that uses extra information is that produced using
Assembly. The Assembly program predicts gene structures from cDNAs and ESTs and is built on top of the Acedb database (4). If there are ESTs or cDNAs that produce different models because of alternative splicing, all models will be generated. This is in contrast to the HMM based method Genomescan that will only produce the most probable model. Assembly is especially focused on aligning ESTs to genomic sequence, for example, it allows for mismatches and frame shifts, and it can use and display the underlying trace data.

Another predicted gene track on the UCSC browser comes from Softbeny (http://www.softbeny.com) and uses a program Fgenesh+, which is based on HMMs and protein similarity but with less emphasis on cDNA/EST data. Fgenesh+ is based on an idea similar to Genomescan and also came out of a group that had previously written abinitio prediction programs.

Gene Prediction Using Other Genomes

With the arrival of the mouse genome sequence in the public domain there is a source of data for gene prediction that is the focus of much gene-prediction research. In theory if you take the genome sequence of two species, the functional parts will be conserved and over time the nonfunctional regions of sequence will slowly diverge. So if you align two systemic regions of the genome, the conserved regions will highlight things like the exons and, possibly, the regulatory binding sites. This has the advantage of not needing a protein or cDNA already in the database to predict genes. In practice the mouse and human sequence have as many matches in introns as in exons, so it is not a trivial problem to sort the coding from the noncoding regions. Additionally not all exons are conserved from mouse to human, so the problem turns from just trivially stringing the matches together to form genes, to a more complicated issue. Several groups have attempted to solve the problem of predicting gene structures by using genomic sequence from more than two organisms. These include Twin scan (6) and Double scan (I. Meyer & R. Durbin, unpublished).

Twin scan is based on the generalized HMM of Genscan and integrates cross- species similarity into the probabilistic model of Genscan. In some ways Twin scan has a more difficult problem to solve than the other methods, as the similar regions between two genomes may fall either in exons or introns or even regulatory binding sites.

Annotation Limitations

At the time of writing, just less than 50 percent of the human genome sequence is available only in draft form. This means that, as well as there being sequencing errors in the draft part of the genome sequence, the fragmented nature of the sequence leads to problems in assembling the genome and subsequently problems in annotating it. The problems in the assembly can lead to genes being fragmented. If for instance, there is misassemble with two regions of sequence being in the wrong order / orientation, it becomes almost impossible for the gene annotation methods to find all the exons for a gene because all the methods rely on exons being on the same strand and in the right order. Also it must not be forgotten that there are still small regions of missing sequence (currently 5% of RefSeq genes cannot be placed on the genome), which can lead to genes being completely absent or only partially annotated. Sometimes if a gene cannot be found in the genome, a close paralog or pseudogene may get annotated as the real gene, since it is the closest thing that can currently be found. This draft nature of much of the sequence makes it very difficult for the automatic methods to differentiate between a pseudogene, and the real gene because they cannot tell whether the difference between errors in
the gene sequence arise from sequencing errors or from real divergence in the genomic sequence. Finally, as all the methods described use, in part, data from the public databases, errors in the database sequences can be propagated onto the genome. This is especially dangerous when building genes using EST sequences that are prone to genomic contamination.

Technology of genome annotation

The process of creating, managing, and displaying genomic information requires the long-term storage of its data and easy programmatic ways to access and update this information. The advantages of using RDMS are as follows: (a) Considerable work has been undertaken in the computer-science industry, providing well-understood, robust systems, in particular with respect to backups and data integrity. (b) Relational databases use a standardized query language (SQL), and all major programming languages have bindings to SQL interfaces, thus facilitating programmatic access. (c) There is a large pool of RDMS- and SQL-savy programmers who can easily be brought into an RDMS-based project. The main implementations are produced by Oracle or Sybase (which are commercial products) or Postgres or MySQL (which are open-source projects). MySQL is not strictly an RDMS, but for practical purposes it can be considered one and is widely used for genome data storage.

A new development in genome viewing is the Distributed Annotation System (DAS) (3). DAS is a protocol that allows a client computer to contact multiple DAS-aware servers and retrieve and integrate genomic annotations. An important feature of DAS is that one does not need to present the entire genome to provide a DAS server—indeed; one can serve up annotations on a small region of the genome. This allows much smaller laboratories, down to modest wet labs with some computer expertise, to run their own DAS servers. The DAS specification has been active for approximately 1.5 years to date, and there are already two stand-alone systems for serving DAS annotations, these being Dazzle from the BioJava project and LDAS, which uses the BioPerl framework. On the client side, the most accessible way to use DAS is using web browsers associated with the main genome projects: Worm base, Gadfly, and Ensembl all offer this capability.

Web Views

Many users of a genome do not have the time or inclination to learn a large complex software system. The genome databases cater to these consumers who are content with web views to the genomic databases. In this review we concentrate on displays provided by the same three main sources of annotation of the human genome previously described (Ensembl, NCBI, and UCSC) because they are the displays we are most used to and because they are the resources most likely to be long lived in their current forms.

In the Ensembl website the front page presents the user with a clickable karyotype, a sequence search interface via either BLAST or SSAHA, and a text search box. The search systems, whether sequence or text, end up on one of two main pages, geneview or contigview, whereas the karyotype ends up only on a contigview page. Contigview shows a slice of the genome, with three levels of contextual information—the position on an idealized karyotype, the general locale of the region of 1 mega base each side, and then a detailed view that is zoomable from 500 base pairs to 1 mega base, which the user controls. The second important page in Ensembl is geneview, which provides a detailed in formation on the intron/exon structure and shows all known identifiers associated with that structure, in particular, Human Genome Nomenclature (HUGO) names, Swissprot, and RefSeq accession numbers. In addition the protein translations of the gene are
compared against InterPro (1), allowing genes with related protein domains to be clustered together.

The UCSC website is focused around a single page that displays a slice of the genome. Either a text search box or the BLAST search page takes the user to this slice, where again tracks of genomic features are shown. The overall paradigm at UCSC is to provide the user with all the information about features on the genomes without imposing too much interpretation about their results. Clicking on the feature of interest provides a small page with UCSC-generated data which generally includes the alignment of the sequence, whereas for contributed tracks a link back to the source website of that feature is provided.

The NCBI website has a two-stage view of the genomic sequence that is heavily coordinated with its Locus Link resource which provides a single point of contact for gene information, with many jumping off points into other areas of NCBI. The map view page provides a view of the genome that is presented as a set of mapping resources on a vertical chromosome. The user can zoom into a region of interest, which will show a scrolled base-pair level view of the genomic sequence.

As with the underlying software there is growing coordination among web sites. Currently the Ensembl and UCSC genome sites are cross-linked at their respective “genome slice” views. This has been a beneficial approach to the user since; each site has strengths in different areas, we are hopeful that this trend of coordination will continue over time, in particular, in the work of other genomes.

Data Resources

As various biologists are accessing specific types of genes or regions of interest through the web, there are a variety of other power users and bioinformaticians who are less interested in web usability and more interested in appropriate slices of the data. Different databases take varying approaches to providing data access. Ensembl probably goes the furthest by providing a fully portable system that can be mirrored, i.e., a standard flat-file style distribution of the data through web-accessible bulk downloads. Ensembl also provides an internet-accessible MySQL host providing full SQL access to the data store to anyone with a MySQL client. UCSC provides a very effective slice of the genome sequence, with server and bulk download of tab-delimited tables that represent the underlying database. One interesting development in this area is the development of advanced query interfaces to the database at Ensembl that allows a user to supply in a web form a query such as “all the validated SNPs within 5 KB of genes with protein kinase domains,” with the result being a formatted Excel spreadsheet. Providing a mixture of web-accessible and bulk-download forms to the users seems likely to be a growing development.

Future Contemplation

With the expected number of new eukaryotic genomes on the horizon is clear that the process of handling genomes will become even better understood, simply by necessity. We expect that the informal sharing of concepts and small components between groups will grow resulting in one or two main sets of components that can be gathered together for genomic annotation. The open-source nature of some of the projects should accelerate this process.

The DAS system democratization of genome annotation (5) shows great promise, with many small laboratories already setting up DAS servers or using the web-accessible DAS upload servers. The presence of DAS prevents genome annotation to be solely in the domain of the large centers and allows both biologists and research bioinformaticians to concentrate on their area of expertise.
Bioinformatika

All sites are providing the ability to integrate “vertically” from the genome up toward proteome and functional aspects of the genome. In addition, many of the systems are being cloned in a “horizontal” manner to provide the analysis and presentation of other genomes. The integration of the entire space, i.e., being able to easily consider queries that, say, efficiently and correctly correlates the quantitative loci traits on rat with the haplotype maps on human are fascinating to consider. It is likely that the next couple of years, with the advent of many new genomes, will force the development of new concepts and potentially new technologies to meet this challenge.

References


*******
AN INSIGHT TO APPLICATION OF .NET PLATFORM IN BIOINFORMATICS

Surya Narayan Rath.

Department of Bioinformatics, Orissa University of Agriculture and Technology, Bhubaneswar-751003, Orissa

Introduction

The completion of Human Genome Project has revolutionized the field of biology and became a milestone for the emergence of a new field like bioinformatics. Bioinformatics is the application of Information Technology to the management of biological information, which is used to gather, store, analyze and integrate biological and genetic informations. The need for bioinformatics capability has been precipitated by the explosion of publicly available informations resulted due to a large scale projects such as various genome sequencing, biomedical, analysis of genomes, proteomics, and biochemical pathways. As a result bioinformatics gave the opportunity to the researchers to access various databases and to use various tools and also to exchange of information through web. Linux operating system is widely accepted due to its high performance quality and open source nature. A large scale of bioinformatics tools has been developed with the need of specialized knowledge and skill set of a trained bioinformatician. But some tools for example Blast become most popular among biologist as its web interface provides as easy interaction between the end user and the application. This indicates the need of a windows application along with a graphical means for user interaction. One of the most interesting examples of this is the recent demonstration by Microsoft research that code found in the Ms Antispyware application would be used to find genetic patterns in HIV.[1] Therefore it is expecting that development of more such tools must enhance the biologist research work. Microsoft .NET Platform is no doubt would be useful in this connection. The .NET Platform provides a graphical user interface to develop a customized and lucrative application which in turns simplifies the interaction between the end user and the application. Visual Studio.NET is the rapid application development (RAD)/ code editor that comes with Microsoft .NET programming language makes designing graphical user interfaces quick and easy.

Discussion

In today's world retrieving information from different publicly available biological database became a necessity for any biologist. Many database searching tool has been developed, which can be categorized as follows:

- Text based database search (e.g., Entrez etc)
- Sequence based database search (e.g. Blast, FASTA etc)
- Motif based database search (e.g. Scan Prosite, Emotif etc)
- Structure based database search (e.g. VAST, DALI etc)

Therefore there is a need of developing efficient searching web tool which can be possible by .NET Framework’s ADO.Net (Disconnected Database Linking Technology) and ASP.NET (Web Hosting Environment) in an easy way. The .NET Framework’s web services strategy handled by XML messaging standard supported by Universal Standard Protocol like HTTP and SOAP (Simple Object Access Protocol) and UDDI (Universal Description Discovery, and Integration) provides a great aids for developing web searching tool. Webservices is just like an application which interacts with different application in a distributed computing environment of the internet. I.e. like other applications Webservices contain business logic (programmable logic) which separates the clients...
from the persistent database. Basing on XML, the Universal language of internet data exchange webservices can communicate across platforms and operating system regardless of the programming language in which applications are developed. [2] SOAP is a XML based messaging technology standardized by World Wide Consortium and communicating between them. UDDI is a public registry, offered at no cost, where one can publish and enquire about the webservices.

The developers using Linux or Mac OS X may feel a bit unhappy with the dependencies of .NET and using Microsoft style configuration (Registry, DCom, Active directory etc). The most vital point of unwillingness to the Microsoft .NET is it's changing of version (layer of libraries) in every couple of years. But the Open Source Project called MONO (supported by Novell) must be good news which brings all .NET goodies as Open Source to Linux, Mac OS X, Solaris etc, on Intel, AMD, PPC, and Sparc. [3] This will enable developers to compile in Linux, run in Mac OS X, compile in windows (Visual Studio on sharp Develop etc) run in Linux etc. But the key benefit of .NET is cross language support for C#, F#, J#, Java, VB, Python (Iron Python), Perl (PerlNet), Ruby, Boo, Pascal etc into one linked application and a nice way to develop applications (GUI designs built into IDE).

Conclusion

Actual data analysis or data managing in .NET is still a question mark .NET has not got wide spread support in Bioinformatics community like Perl. Python and Java as these are freely available. As a result certain gap is maintained between Bioinformatics and .NET platform.

Web links


COMPANY PROFILE

Jubilant Biosys Ltd. (JBL), a Bangalore-based company, is focused on providing innovative informatics and structure-directed drug discovery solutions to help accelerate the process of global pharmaceutical drug discovery.

The mission is to develop an environment for efficient and sustained innovation for the collaborators and employees in addition to protecting and respecting intellectual property. With superior technologies, world class talent and cutting-edge research, JBL is committed to delivering high quality, cost effective solutions to aid the customers in the process of drug discovery.

JBL, established in 2001, has graduated from being a leading provider of discovery knowledgebase to an integrated collaborator in global pharmaceutical early lead discovery.

At JBL, over 600 scientists experienced in chemistry, biology, biochemistry, pharmacology and clinical areas, and domain-specific IT experts, strive to deliver cost-effective innovation, driven by speed and efficiency.

*******

Jubilant Biosys Ltd. (JBL), a Bangalore-based company, is focused on providing innovative informatics and structure-directed drug discovery solutions to help accelerate the process of global pharmaceutical drug discovery.

The mission is to develop an environment for efficient and sustained innovation for the collaborators and employees in addition to protecting and respecting intellectual property. With superior technologies, world class talent and cutting-edge research, JBL is committed to delivering high quality, cost effective solutions to aid the customers in the process of drug discovery.

JBL, established in 2001, has graduated from being a leading provider of discovery knowledgebase to an integrated collaborator in global pharmaceutical early lead discovery.

At JBL, over 600 scientists experienced in chemistry, biology, biochemistry, pharmacology and clinical areas, and domain-specific IT experts, strive to deliver cost-effective innovation, driven by speed and efficiency.

*******
In addition to their usual academic responsibilities, NISTians continued with their multifarious activities in different areas of Science & Technology. Detailed description of each one of those are beyond the scope of this section. Instead an attempt will be made to provide a snapshot of the activities (over the last 3 months) in areas of Biotechnology/Bioinformatics and allied areas like Soft Computing, Intelligent Systems etc. which have direct relevance to them. Persons desirous of possessing more details should visit www.nist.edu.

1. FACULTY DEVELOPMENT PROGRAM :

An AICTE Sponsored Faculty Development Program on Modelling, Simulation & Optimization in Science & Engineering was held at NIST during July 02 - July 14, 2007. The same evoked considerable enthusiasm resulting in receipt of applications well above the specified limit (as stipulated by AICTE) from all over India. Faculties/Researchers from various reputed institutions, apart from those of NIST itself, delivered lectures on various aspects of Modelling, Simulation & Optimization. The Director of NIST, Prof. Sangram Mudali in his inaugural speech, emphasised the importance of Modelling, Simulation & Optimization in different aspects of Science & Engineering. He also indicated how he developed and cultivated his fascination for these since his student days. Prof. A. Mukherjee of Indian Institute of Technology, Kharagpur, graced the inaugural session as the Chief Guest. Dr. A.K. Padhy, Course Co-ordinator, B.Tech, NIST conducted the proceedings of the same. At the valedictory session the participants were awarded the certificate of participation by Prof. Sangram Mudali, Mrs. Geetika Mudali, Placement Director, NIST & Prof. Ganapati Panda, National Institute of Technology, Rourkela, Orissa.

2) EDITORIAL RESPONSIBILITIES :

(i) Editor-In-Chief

Prof. Suash Deb has been appointed as the Editor-in-Chief of a forthcoming, newly launched publication - International Journal of Soft Computing & Bioinformatics. Reputed researchers from India (IITs, ISI, IIM & Others) as well as those from USA (Harvard University, New Jersey Institute of Technology etc.), Australia (Monash University), New Zealand (Auckland University of Technology), S. Korea, Japan, West Indies, Poland, Spain etc. have already joined the International Editorial Board of the same.

The Journal is expected to commence its journey from Nov-Dec’07. Anybody wishing to subscribe to the journal and/or intend to submit manuscript for possible publication in this journal is advised to get in touch with Prof. Suash Deb.

(ii) Regional Editor

Prof. Suash Deb has been appointed Regional (India & Subcontinent) Editor of the journal - Neural Computing & Applications (a Springer publication). Personnel from the above regions desirous of submission of articles in various aspects of Neural Networks and/or making subscription should contact Prof. Suash Deb.
**ALUMNIS’ DELIGHT AT INFOSYS**

Date : Tue, 19 Jun 2007 17:12:45 +0530  
From : “gayatri g” <gayatri.345@gmail.com>  
To : “suash deb” <suashdeb@yahoo.com>  
Subject :  
Hello Sir,  
How are you ? I am fine. At present, undergoing training at Infosys, Mysore. All of us from CS and IT are in fast track. rest all are in long cycle. Training is going fine.  
The ground rules of being systematic and disciplined, coming to classes on time, which you have taught us at NIST is helping us a lot! Has the college reopened?  
I saw the last two issues of Bioinformatika at Infosys library here. It feels very nice to see our Bioinformatika sharing space with other journals across the world.  
Am waiting eagerly for the next issue.  
Plz mail me.  
Regards,  
Gayatri  
Infosys Technologies Limited.

Date : Thu, 21 Jun 2007 19:38:07 +0530  
From : “Swastik Choudhury” <swastikchoudhury@gmail.com>  
To : suashdeb@yahoo.com  
Subject : Hello sir  
Respected Sir,  
How are you? We all are fine here. Last week as I was wondering through our INFOSYS library I came across both the issues of our BIOINFORMATIKA magazine.  
It gave me immense pleasure and surprise to feel that I was a part of it and was able to contribute something back to my college under your able guidance.  
How is Jyoti sir? I miss his jokes.  
How is college going on and the work on BIOINFORMATIKA?  
I will be missing the fun I got while working for BIOINFORMATIKA.  
Looking forward to its next issue in INFOSYS library.  
Thanks and Regards,  
Yours obediently,  
Swastik Choudhury,  
Software Engineer, Infosys Technologies Limited.

Date : Thu, 21 Jun 2007 18:53:49 +0530  
From : “Jyoti Prakash Dash” <jyotiprakashdash342@gmail.com>  
To : suashdeb@yahoo.com  
Subject : Bioinformatika at Infosys Library  
Respected Sir,  
It is really nice to see the NIST Publication of Bioinformatika making its way to the Infosys Library alongside many famous journals and Magazines.  
It makes me feel real proud to see the College Club Magazine standing as an informative journal providing vital information on day-to-day Bioinformatics events.  
This will always prove as a ladder for our college to be recognised in the global arena.  
Best of Luck to the entire team guided by Suash Deb sir for their work.I hope that the team will produce more publications that will be famed at different universities.  
Regards,  
Jyoti Prakash Dash  
Infosys Technologies Limited.
Date : Thu, 21 Jun 2007 18:52:14 +0530  
From : “chinmayanand choudhury”  
<chinu.2k17@gmail.com>  
To : suashdeb@yahoo.com  
Subject : Excellence of Bioinformatika @ infosys

Respected Sir,

Good Evening!

I am delighted to see copies of Bioinformatika at infosys library.

Because of its excellence it is now being recognized at industry as well.

I wish, this magazine will explore various areas of Biotechnology and Bioinformatics in the coming issues.

My best wishes to the entire bioinformatika team

Missing you all.

Good bye

With regards,

Chinmayanand Choudhury,
Software Engineer, Infosys Technologies Limited

-----------

Date : Thu, 21 Jun 2007 06:20:47 -0700 (PDT)  
From : “geeta devi” <geeta_294@yahoo.com>  
To : suashdeb@yahoo.com  
Subject : Hello Sir

Respected Sir,

It’s really very nice to see the NIST Bioinformatika magazine here at Infosys alongwith other famous magazines. I really feel proud to see many people at Infosys library reading our college journal.

Regards,

Geeta
Infosys Tecnologies Limited

-----------

Date : Mon, 13 Aug 2007 22:59:37-0700 (PDT)  
From : “pragyan patnaik”  
<pragyan_patnaik43@yahoo.com>  
To : suashdeb@yahoo.com  
Subject : Excellence of Bioinformatika @ infosys

Respected Sir,

I am delighted to see copies of Bioinformatika at infosys library.

Because of its excellence it is now being recognized at industry as well.

I wish, this magazine will explore various areas of Biotechnology and Bioinformatics in the coming issues.

My best wishes to the entire bioinformatika team

Missing you all.

Good bye

With regards,

Pragyan Prabodhita Patnaik (7th sem, IT)

-----------

Date : Mon, 13 Aug 2007 07:18:54+0100 (BST)  
From : “Monalisa Sahu”  
<monalisa_sahu_10@yahoo.co.in>  
To : suashdeb@yahoo.com  
Subject : Excellence of Bioinformatika @ infosys

Respected Sir,

Let me express my thanks to you for giving us a golden opportunity to gain some knowledge about Bioinformatics in 6th semester and also continue with Soft Computing in the 7th semester. Initially I was afraid of Bioinformatics. But when I was acquainted with your teaching style, I really felt very much comfortable with the subject. Subsequently, the publication of Bioinformatika made me generate extra interest in this emerging topic. As you are very loyal towards your duty and quiet disciplined in nature I feel the class environment extremely educative. You are so industrious that the success will always remain with you.

With regards,

Monalisa Sahu (7th semester, IT)
The Editorial Board of Bioinformatika is undertaking a formal evaluation of its every issue. The aim is to make the future issues more interesting, more informative & more useful for our readers. Towards that we solicit your cooperation in completing the following questionnaire and return the same, at your earliest convenience to either by e mail or post to:

Prof. Suash Deb, Editor-in-Chief, Bioinformatika

QUESTIONNAIRE:

1. Please indicate your assessment of the Quality of the Publication w.r.t.
   - Presentation / Format: 4 3 2 1
   - Readability: 4 3 2 1
   - Timeliness of Publication: 4 3 2 1
   - Coverage of Subject Matter: 4 3 2 1
   - Analytical Rigour: 4 3 2 1
   - Overall Quality: 4 3 2 1

2. How useful is the publication for your work?
   - Provision of Information: 4 3 2 1
   - Clarity of Information: 4 3 2 1
   - Clarity of Findings: 4 3 2 1
   - Message Suggestions: 4 3 2 1
   - Overall Usefulness: 4 3 2 1

3. Please give examples how this publication has contributed to your work:

4. Your Suggestions For Improving The Publication:

5. Your Background (Without This Information, Your Feedback Would be Treated as Inconsequential):
   - Name:
   - Title/Position:
   - Institution/Organization:
   - Office Address:

Please use additional sheets of paper; if necessary. Thank you for your time for completing this questionnaire.
"I am sure in the coming days Bioinformatika will play a very cogent role towards popularizing Bioinformatics in our country."

- Prof. S. K. Sanyal
(Hon’ble V.C.- Jadavpur University, Kolkata)