Plant Tissue Cult. 11(2): 195-208, 2001 (December)

PTC

Internet Resources for Molecular Biology : A Primer for Scientists from Least Developed Countries

Hemayet Ullah, Gregory Clark[1] and A. S. Islam[1]

Biology Department, University of North Carolina at Chapel Hill, NC 27599, USA. E-maill: ullah@email.unc.edu

Never before in the history of mankind, has so much scientific information been compiled, collated, annotated and above all made available to the general public through the Internet. The exponential growth characterizing the steep rise in the volume of such information as seen in the last few years has been illustrated in Fig. 1. One such source of information is PubMed (http://www.ncbi.nlm.nih.gov/Entrez), which covers citations of almost all biological and Medical sciences with links directly to online articles. PubMed allows one, without any cost, access to nearly 11 million such citations. The variety of available literature via the Internet represent several facets of biological science such as (a) 11 million titles with abstracts of all important articles published in international journals, (b) the DNA- and protein sequences deposited to GenBank or the EMBL database, (c) the maps of more than 30,000 human genes (http://www.ncbi.nlm.nih.gov/genemap99), and (d) ~16300 different kinds of 3-dimensional protein structures (http://www.rcsb.org/pdb/) in the Protein Data Bank (PDB). In less than three years, the PDB has registered more than a twofold-increase in the number of protein structures, from ~7500 to ~16300.

How to Navigate such a Vast Wealth of Information: At this point in time, question that arises in the mind of a user is how to navigate such a vast wealth of information and retrieve the relevant data. In order to dig out the necessary information from the Internet, it is essential that the concerned scientist acquires workable knowledge in bioinformatics, a rapidly growing science combining computer- and biological science. Bioinformatics has proved to be a very powerful tool in the hands of molecular biologists to 'mine' important data with precision, particularly in the wake of recent hardware and software advances.

---

[1]MCDB, University of Texas, TX 78713, USA.

Globalization of the Internet Offers Unlimited Opportunities to the Third World Scientists: The potential user is confronted with a number of difficulties. For instance, the volume of DNA sequences available in various databases is by far greater than the number of publications, implying that scientists have not been able to keep pace with the analysis of the accumulated data. There is no question that this gap will soon be filled with the availability of more sophisticated software enabling the scientists to do this seemingly difficult task. However, the globalization of the Internet and free access to it offers a great hope to the scientists from the third world countries in general and least developed countries in particular. In spite of the fund constraints, scientists from this part of the world, once they train themselves in the essentials of bioinformatics, will be able to utilize such raw data to figure

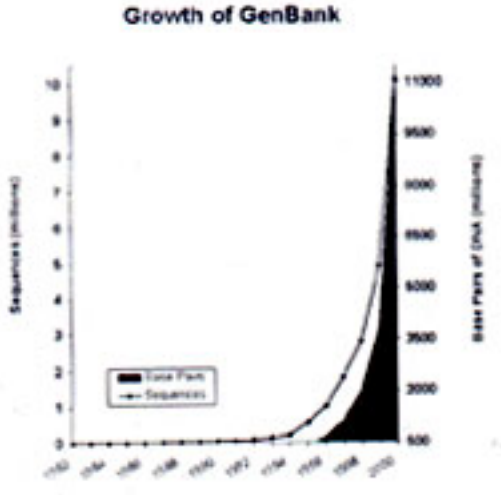

Fig. 1. Exponential growth in the DNA sequences deposited in Genbank. Source: www.ncbi.nlm.nih.gov

out functions of specific genes of interest. Enough resources such as data mining tools, research databases, online journal abstracts as well as training courses are available on the Internet for this purpose. In other words, without the commitment of sophisticated technical and financial resources, it is now possible for a scientist with Internet access, regardless whether he is from a developed or a developing country, to make an intelligible interpretation of the biological data at his disposal. Utilizing bioinformatic tools he may be able to 'guess' the 'putative' function of previously uncharacterized genes (sequences, clones). These 'guesses' can

be substantiated by designing experiments with 'low cost' equipment(s) which scientists from this part of the world may be able to afford from their own research funds .The underlying benefits of this kind of research, utilizing information from the Internet, would be considerable, as it would contribute to the world pool of knowledge. For instance, this approach could be applied to deciphering the gene functions of tropical and semitropical plants.

This article has been developed to give a brief description of all of these resources and to explain, in detail, how these Internet sources may be utilized to do cutting edge research while spending the least amount of available financial resources. The examples will show how one can assign a putative function to an unknown gene and test its putative function(s). This article attempts to provide a pragmatic hands-on approach to solving problems at the interface of experimental and theoretical molecular biology.

A researcher must be clear as to the kind of databases to be used to meet desired goals. For example, a plant biologist may be interested in finding genes that regulate mitosis in a certain tropical plant species. By surveying the literature, the investigator should be able to retrieve the information that the B-type cyclin commits cells to undergo mitosis from the G2 phase of the cell-cycle in yeast, man, and mammals. Equipped with this information, the next step would be to carry out a detailed search to find out whether cyclins are present in any plant species, or their close relatives. The fact that there are only few functional pieces of evidence on the occurrence of cyclins in plants as such, makes the task more difficult. Fortunately, like a dictionary, the entire base sequence of the *Arabidopsis* genome is now available. He can look for the cyclin gene, scanning the entire map containing the *Arabidopsis* base sequence utilizing BLAST (see below); and in the event of finding analogous (similar) sequence in the *Arabidopsis* genome he may proceed to present functional data to support that the cyclin protein is indeed encoded by the gene initiating mitosis in plants. Functional data can be presented by obtaining seed from a plant (from another specific database on the Internet) in which the identified gene has been deleted ('knock-out'). However, note here at present, gene knock-out plants are available only from *Arabidopsis*. If the identified gene indeed encodes for a cyclin that promotes mitosis, the 'knock-out' plant should have a lower rate of mitosis and as such the mutant individual would be characterized by fewer cells. Below, a real-life situation is presented to show how a researcher should go about to

find homology of the gene of interest to similar genes in other organisms: plant or animal. The example used here is the G-protein coupled signal pathway – one of the main pathways that animal cells use to transduce signals for vision, smell, neurotransmitters, hormones, and other critical processes.

G-Protein from *Arabidopsis*: Considering that animals share many of the same signal transduction pathways with plants raises the question, how plants use these pathways, and how they are different from the animal cells. Dr. Ma and his associates provided a partial answer to this question more than a decade ago (Ma *et al.*, 1990, Weiss *et al.* 1994). They cloned genes encoding G-proteins similar to those reported earlier in animals from *Arabidopsis*. However, one of the main upstream components of the pathway, namely, the G-protein coupled receptor (GPCR) was not identified in the above report. Josefsson and Rask (1997) discovered this component in plants. Since animal cells use different GPCRs to couple different signaling molecules, it becomes obvious that plant molecular biologists would try to find additional GPCRs from plant genome, particularly in the wake of the *Arabidopsis* genome sequencing.

- The first logical step in this direction will be to go to the 'mother of all biological databases', namely, the National Center for Biotechnolgical Information (NCBI at http://www.ncbi.nlm.nih.gov)
- Search the database (GenBank) with a text based query *"Arabidopsis* G-protein coupled receptor". The search would return 21 sequences that have some relationship with *Arabidopsis* GPCR.

After a careful analysis of the reported sequences, only the one containing the full-length of *Arabidopsis* GPCR mRNA is selected by clicking the accession number, AY054650. At this stage it is to be ascertained whether any other worker has reported this particular sequence and its putative function. If not, the investigator may proceed to characterize this basically 'untouched' sequence together with the object of determining its function. The other 'hits' are discarded because most of these sequences are partial mRNAs or those from other plants that show similarity to the *Arabidopsis* GPCR. However, even if the investigator finds a gene that is annotated as *Arabidopsis* GPCR, how would the user go about to verify whether this gene is indeed among 'true' GPCRs? To perform this task, the researcher needs the use of a "BLAST" search.

BLAST (Basic Local Alignment Search Tool) search : In order to verify that the sequence earmarked by the investigator indeed encodes for a 'true' GPCR protein, the user needs to find out if it has a similarity with sequences of other known GPCRs. Therefore, searching protein/nucleotide databases with the *Arabidopsis* GPCR sequence as a query is the next step. Sequence comparison process should be done with extreme caution! Interpretation of results and the design of search parameters should be done exercising the utmost care. BLAST search engine found at the NCBI's web page is an excellent software resource to search for similar sequences (http://www.ncbi.nlm.nih.gov/BLAST)

Depending on the query sequence, the investigator should choose a particular program to run the query. For example, if the query is about a DNA sequence and is meant to seek identical DNA sequences, the researcher should choose the 'blastn' program. The following table explains the function of different programs.

Program	Query	Database	Comaprison	Use
blastn	DNA	DNA	DNA level	Find identical DNA sequences
blastp	Protein	Protein	Protein level	Find similar proteins
blastx	DNA	Protein	Protein level	Translate DNA sequence in protein and find similar proteins
tblastn	Protein	DNA	Protein level	Database translated in proteins and find similar proteins
tblastx	DNA	DNA	Protein level	Both query and database translated (most time consuming but probability of finding similar sequences is higher

For more advanced explanations, readers are directed to the following two web pages: http://www.ncbi.nlm.nih.gov /blast/html/ blastcgihelp.html#protein_databases. http://www.ncbi.nlm.nih.gov/ Education/ BLA ST info/ Blast_setup.html#Plan.

The BLAST pages offer several different databases allowing researchers to do the specific search they are interested in. Some of these databases are: SwissProt, PDB and Kabat. They represent compilations of protein/nucleotides sequences outside of NCBI. However, for all practical reasons, the researcher is advised to select the non-redundant (nr) database. The advantage of the nr database is that it does not include redundant sequences and it stores only one copy of a particular sequence compiled from different databases, thus removing any redundant sequences for a particular gene.

To do a BLAST search, the investigator copies the protein sequence for the *Arabidopsis* reported GPCR that the user found earlier by doing a text-based query (accession# AY054650) and pastes the sequence into the search page of 'blastp' whose link is found at http://www.ncbi.nlm.nih.gov/BLAST. The results are returned with a list of genes that show similarity to the investigator's query sequence and some statistical parameters to show the confidence level in the calls. The most important assessment for confidence in the reported list comes from the E-value (Expect value). This value represents the number of distinct alignments with similar scores that could have been selected purely by chance. For example, an E value of 1 assigned to a hit can be interpreted as meaning that in a database of the current size one might expect to see 1 match with a similar score simply by chance. Therefore, the smaller the E-value, the better is the confidence in the reported similarities being real (not by chance). In our list, all top scoring 'hits' have E-values close to zero and they all came from GPCR-related sequences, thereby increasing the confidence that the query sequence could be a 'real' GPCR. A hit with potato GPCR is shown below with an e-value of e-148. However, there is another group of proteins- lantibiotic synthetase C-like protein 2, which also shows a high similarity with significant e-values. Thus, at first glance, AY054650 looks like a true GPCR; however, without further 'functional' evidence, the result should be considered tentative.

>gi|8489877|gb|AAF75794.1|AF272710_1 (AF272710) putative 7-transmembrane G-protein-coupled receptor [Solanum chacoense]

Length = 424

Score = 525 bits (1353), **Expect = e-148**

Identities = 261/433 (60%), Positives = 320/433 (73%), Gaps = 12/433 (2%)

```
Query: 2.   SSSVDFVREQGRCGDDGNGAGETVKNGEIDHLLSEPSAPT
             IS LPTESFLRAATLLKNQVV 61 SS V Q  DDGN +T ++ +P+A++
             S +E+FL+AA  LK+QVV

Sbjct: 3.   SSVVQLTASQKNNSDDGNERVDTDQH-----HQPTAHSV SHT
             SETFLQAAISLKDQVV  55

Query: 62.  EATWKGGVEALASGSGPVLDPTVYTGLLGTAFTCL
             KSYEVTRNHQDLLTCAEIIDTCANV 121 E TWK +G V DPT
             +YTGLLGTAFTCL+SYE T + +DL  C+EI+D CA++

Sbjct: 56.  EMTWKEN----GRSAGSVTDPTMYTGLLGTAFTCLRSYEA
             TGDRKDLELCSEIVDACADL 111
```

Fig. 2. Alignment of reported similarity with the query sequence of *Arabidopsis* putative 'GPCR' with a potato GPCR. On the alignment lines, + represents similar (with same properties like acidic or basic) amino acids.

Though the alignment results are statistically significant, the analysis needs further refinement to have a confidence in assigning the sequence to a putative GPCR. There are numerous options available to further analyze the sequence. For example, one may use a database that only deals with GPCRs (http://www.gpcr.org/7tm/). Consulting this database will provide the investigator with relevant information such as structures, locations, gene expression patterns, functions of different classes of GPCRs. As the database also contains sequences of all known GPCRs, an alternative program called Clustal W can be used in order to obtain an alignment of multiple sequences. The program is available at: http://dot.imgen.bcm.tmc.edu:9331/multi-align/multi-align.html. This allows the investigator to compare the putative GPCR sequence to all known GPCRs and to determine whether the putative GPCR has the critically conserved residues present in other known GPCRs.

It is well-known that GPCRs are proteins located in the plasma membrane with 7 transmembrane domains. Therefore, the obvious question would be whether the putative GPCR has this characteristic 7-pass domain structure and is located in the membrane itself. There are several websites that use different algorithms to predict the location of a sequence of this kind. A quick check at the http://www.sbc.su.se/~miklos/DAS/ reveals with confidence that the putative GPCR sequence is indeed predicted to be in the membrane and that the sequence has 7 potential transmembrane domains.

Gene knock-out: At this point, the investigator develops a positive idea that the sequence may represent a 'true' GPCR. The next step is to obtain a mutant plant with the gene of interest knocked out. Gene 'knock-out' is a powerful technique to determine the function of a gene product. If the knock-out plant shows an abnormal phenotype, for example, if it does not respond to a growth hormone, the investigator may safely conclude that the gene of interest is needed to perceive said hormone. In *Arabidopsis*, a relatively new technique has been developed to generate mutant plants in which insertion of foreign DNA into the gene of interest has become a routine operation (Krysan *et al.*, 1999). Insertion of a piece of foreign DNA (called T-DNA) into a gene of interest effectively 'destroys' the gene from the genome. These random insertions have created ~100,000 mutant plant lines, enabling the investigator to look for a knock-out mutant plant containing the gene of interest. There are several databases on the Internet where one can search to see whether there is an insertion in one's gene of interest. Two of them are at

http://signal.salk.edu/cgi-bin/tdnaexpress and http://www.tmri.org/pages/_collaborations/garlic_files/GarlicAnalysis.html.

While searching the Salk Institute's T-DNA insertion lines, the first author could not find a 'knock-out' mutant plant containing a T-DNA insertion in the GPCR sequence. However, the Garlic database has ~90,000 lines and submitting the GPCR sequence to that databases results in finding a plant in which the GPCR gene has been 'knocked-out'. Failure to find any insertions does not mean that this sequence for G-Protein gene is lacking in *Arabidopsis*; it means that the investigator needs to wait until sequencing of all the insertion lines is completed with an entry of this particular sequence in the database.

Once the investigator locates such an individual in the database, the next step will be to procure sufficient seed from either of the sources mentioned above and raise the population containing the knock-out gene in the laboratory. Mutants should be grown under different growth conditions, different types of stress and with or without supplements of hormone(s) followed by scoring the population for any altered phenotype. The mutant phenotype will be characterized by the lack of expression of the trait conditioned by the "knock-out" gene. Conspicuous absence of a morphological or biochemical character will help the investigator associate the "knock-out" gene with that particular trait. While scoring the population for a trait, a point to remember is whether the mutants carry "knock-out" genes that are essential for vital metabolic processes such as respiration, photosynthesis, or reproduction; homozygous knock-out mutants of this class may be lethal.

There is a wealth of biological information available on the Internet. Pedro's Biological Database was once 'the' database for all molecular biologists looking for relevant information. This database, in spite of the fact that it has not been updated since 1996, is still very useful and user-friendly for an investigator to locate various sites for molecular data analysis. Pedro's database is organized into 5 different sections as follows: Part 1: Molecular Biology Search and Analysis; Part 2: Bibliographic Text and WWW Searches; Part 3: Guides, Tutorials and Help Tools; Part 4: Bio/Chemical Journals and Newsletters; Part 5: List of Pedro's Biological Resources. The database lists a substantial number of easy-to-use molecular biology tools (simply cut and paste the sequence of the gene of your interest into the space allotted for pasting and click on 'submit' button). Here is a list of some functions one can do utilizing the programs listed in the Pedro's database. Translation of DNA

sequence into protein sequence with all possible orientations (6 possible orientations); Protein Identification using Amino Acid Composition; Compare the Amino Acid Composition of one Protein with the other; navigate database of Highly Conserved Regions in Proteins; Prediction of Coiled Coil Regions in Protein Sequences; navigation of Enzyme and Metabolic Pathway (EMP) Database; predict Gene Structure, Internal Exons and Splicing Sites in DNA, Exon-Exon Junction in cDNA, and designing primers for PCR.

Currently, the most comprehensive data analysis database is maintained by the website of Pasteur Institute of France (http://www.pasteur.fr/recherche/BNB/bnb-en.html). The web page lists 1344 online analysis tools, 1427 different databases, 1044 courses and tutorials, 1098 different softwares. One can search the whole database by resource type, by organism, or by different biological domains (biochemistry, genomic, genetic, structure, biocomputing, evolution etc.) or even by country resources. Readers will find a list of several references that will guide them to navigate the Internet for garnering a wealth of biological information.

The steps described above may be summarized as follows:

1. Open up NCBI home page (www.ncbi.nlm.nih.gov)
2. Go to drop down menu at the top left hand and select "protein".
3. In the same page write in the box provided at the top within double quotes "*Arabidopsis* G protein coupled receptor". Without double quotes two many references are pulled up because the computer searches for all the three words independently.
4. From the pulled up page, choose # 9 where the sequence of 401 amino acids sequence for the putative G-protein is shown. COPY THE SEQUENCE.
5. Go to Pedro's biomolecular research tool page http://www.public.iastate.edu/~pedro/rt_1.html and click TMpred.
6. If it does not work attempt the Canadian web site ExPasy Proteomic tools: http://ca.expasy.org/tools/#transmem.html.
7. Move your cursor to nearly the bottom of the page and hit DAS in the section under, "Transmembrane regions detection". You will see "DAS" -Transmembrane Prediction server. In the blank space captioned (Use protein sequence only!), paste the amino acid sequence of the gene product of your interest.
8. Go to the Salk Institute Genomic Analysis page: http://signal.salk.edu and hit T-DNA caption on the left hand side of the *Arabidopsis* figure.
9. In the T-DNA page paste the amino acid sequence, the protein product of the G protein of your interest and SUBMIT. It will yield no results indicating that in their population, there are no individual plants with T-DNA inserted in the searcher's G-protein. However, submitting the same sequence to the another T-DNA insertion population webpage (http://www.tmri.org/pages/collaborations/garlic_flies/Garlic_Analysis.html) will allow one to get a T-DNA insertion Arabidopsis plant line.

PCR amplification of a gene of interest : Let us deal with one more example. If a scientist is looking for a gene of interest and that the gene is not specifically annotated in the *Arabidopsis* genome, a more difficult approach must be used. Examples of this case might be a gene such as the inositol trisphosphate receptor or integrin-like receptor for which there is experimental evidence for plant functional homologs but no known gene sequence in plants thus far. In this type of analysis, the researcher would attempt to find in the literature protein sequence data for functional domains within these proteins that may be highly conserved across diverse organisms.

The researcher could also do protein alignments of all the animal versions of the protein in question and look for peptide sequences that are conserved in all of the known examples of this protein. Once some conserved peptide sequences from the gene of interest have been found tblastn (NCBI website as mentioned above) can be used to survey the *Arabidopsis* genome for potential hits. Alternatively, the researcher could design redundant primers based on the conserved animal peptide sequences from the user's search (at least 6-7 amino acids long) and do PCR with genomic or cDNA libraries from the plant species of interest in an attempt to amplify part of the gene. This approach would also be useful in trying to obtain a gene in a plant species, which has no homolog in *Arabidopsis*.

In designing primers there are many different programs that can be purchased and used if the sequence of the gene is available, as is the case with many *Arabidopsis* genes. There is also free access to a Web based program that is one of the best for designing primers when a researcher has the nucleotide sequence. This website is provided by the Center for Genome Research at the Whitehead Institute for Biomedical Research of Cambridge, Massachusetts, USA (http://www-genome.wi.mit.edu/). Once at this website, the researcher scrolls down to the sub-heading called Genome Center Software and clicks on WWW Primer Picking (Primer 3). At this point the scientist can cut and paste the gene sequence that the user wants to design primers for and the program will pick several sets of primers. The program works nicely without any changes to the default settings, but there is one setting located under the heading of global parameters titled "product size range" that the researcher may want to adjust. DNA products in the range of 500 bp to 2000 bp are easiest to work with. The template to be used in PCR is most often a first strand cDNA made from isolated RNA/mRNA following a reverse

transcription reaction, but genomic DNA can also be used if the researcher is interested in obtaining the genomic version of the gene including introns.

In designing redundant primers to be used to amplify a putative gene in the absence of sequence data for that gene, there are some basic guidelines to follow. The two primers should be from conserved amino acid sequences that are located at least 30 amino acids apart and no more than 700 amino acids apart. The amino acids to be used for the PCR primer cannot be too redundant, so multiply the number of triplet codons for each amino acid to calculate redundancy (<250 is good) (refer to Fig. 3). The last two amino acids for the forward primer (left primer; N-terminal) and the first two amino acids for the reverse primer (right primer; C-terminal) need to be low in redundancy and highly conserved

A	alanine	GCT; GCC; GCA; GCG
C	cysteine	TGT; TGC
D	aspartic acid	GAT; GAC
E	glutamice acid	GAA; GAG
F	phenylalanine	TTT; TTC
G	glycine	GGT; GGC; GGA; GGG
H	histidine	CAT; CAC
I	isoleucine	ATT; ATC; ATA
K	lysine	AAA; AAG
L	leucine	TTA; TTG; CTT; CTC; CTA; CTG
M	methionine	ATG
N	asparagine	AAT; AAC
P	proline	CCT; CCC; CCA; CCG
Q	glutamine	CAA; CAG
R	arginine	CGT; CGC; CGA; CGG; AGA; AGG
S	serine	AGT; AGC; TGA; TCT; TCG; TCC
T	threonine	GTT; GTC; GTA; GTG
W	tryptophan	TGG
Y	tyrosinine	TAT; TAC
STOP		TAA; TAG; TGA

Fig 3. The genetic (triplet) code. The three bases of an mRNA are designated here as the first second and third base, reading in the 5' to 3' direction along the mRNA. Three of the 64 codons function as "stop" signals.

because these nucleotides are the most important in annealing with the template. In cases, where there are more than four nucleotide triplets coding for a single amino acid and the triplets differ in the third position, the researcher can make a primer with inosine (I) at the third position. For

example, with the amino acid glycine a primer can be made using GGI for glycine because all four nucleotide triplets that code for glycine start with GG and by using an inosine in the third position all four possibilities will be covered (refer to Fig. 3). After designing the redundant primer the researcher also needs to calculate the theoretical melting temperature of the primers. Each G and C base can be counted as four degrees and each A and T base can be counted as two degrees and an I base does not count. A final melting temperature for each primer ideally should be above 50 degrees; however a researcher can attempt to use primers with lower theoretical melting temperatures. The researcher needs to keep in mind that the melting temperatures of the primers to be used dictate the annealing temperature to be used in PCR. The template for the PCR reaction can be either first strand cDNA or genomic DNA as discussed above. However, in using redundant primers in search of a putative gene the researcher should use both templates, as the genomic DNA template is more likely to contain the gene of interest. Of course a product amplified using genomic DNA as the template may also include introns.

Currently, the most comprehensive database is that displayed in the website of Pasteur Institute of France (http://www.pasteur.fr/recherche/BNB/bnb-en.html). The web page lists 1344 online analysis tools, 1427 different databases, 1044 courses and tutorials, 1098 different softwares. One can search the whole database by resource type, by organism, or by different biological domain (biochemistry, genomic, genetic, structure, biocomputing, evolution etc.) or even by country resources. Readers will find a list of several references that will guide them to navigate the Internet for garnering a wealth of biological information.

Conclusion : Considering that developing countries have limited resources and have their priorities pinned down mainly to improvement of economically important crops, they need to develop facilities which would yield the maximum cost/benefit ratio. The best way for developing countries to do it is to establish sequence databases for regionally important crops like jute, indigenous cotton, Pennisetum, Amaranths, jackfruit, mango etc and at the same time derive maximum possible benefits from the existing worldwide Internet services providing sequence and related databases.

The first step towards the fulfillment of this objective, need to be directed to establish databases with EST (Expressed Sequence Tags) sequences. The EST sequences are derived from sequencing cDNAs

randomly. It is not feasible costwise for a least developed country to sequence the whole genome of a particular crop plant. However, creation and maintenance of such databases like EST will be less costly compared to highly cost-intensive databases, that are available on the Internet run by international organizations, universities and research institutions of the developed countries . World bodies such as FAO, UNESCO and ICGEB (The International Centre for Genetic Engineering and Biotechnology) may be approached to generate such databases, apportioning one crop, unique to the country. For instance, Bangladesh may be given the responsibility to create and maintain a database containing base sequences of the two jute species, *Corchorus capsularis* and *C. olitorius*. All sequence databases, that are now in operation, started with EST sequences and they were created from random cDNA clones. Existing facilities, such as DNA sequencing apparatus available in most of the developing countries manned by adequately qualified technicians would enable local molecular biologists to jump-start an EST database. Once the DNA sequences of a crop are available, investigators can always use the existing Internet resources to analyze their data.

For example, the lignin content of fibers reduces the quality of natural fibers. Genes, which reduce the lignin content of fibres, have been reported from many organisms such as pine, *Arabidopsis*. *CCR1* of *Arabidopsis* that encodes for cinnamoyl CoA reductase has been identified. This enzyme catalyzes the reduction of cinnamoyl CoA esters into their corresponding aldehydes, i.e. a step dedicated to the lignin biosynthetic pathway. As the sequence of this gene is known, creation of an EST database of jute will enable the investigator to find a cDNA segment in it with a similar sequence as that of the CCR1. Since under Indian subcontinent conditions, establishment of T-DNA insertion lines in jute would be a far cry, the next best way to disrupt this gene will be to use antisense technology in order to reduce the lignin content of jute fibres. Accomplishment of this seemingly difficult task will make the jute crop a tough competitor of cotton, much to the advantage of countries such as Bangladesh where the production of high staple cotton varieties falls far short of the country's requirement to keep its textile mills running. Before a molecular breeder embarks upon a program of introducing a gene for a particular trait such as breeding for salt- or drought tolerance, the preliminary work should comprise navigating the Internet using the sources described above. Such a navigation would provide the breeder valuable clues regarding the structure of the gene of interest, the structure

of its promoter and the kind of regulatory gene that control its expression. The breeder may then negotiate with the concerned group of scientists to procure the gene of interest including the promoter and the regulator that drive its expression. State of the art knowledge, adequate facility of the lab and successful import of the gene of interest will help the breeder reach the target faster, quicker and with precision not hitherto known to biologists.

Acknowledgement

Our grateful thanks are due to Len Krall and Zack Nimchuck of Biology department, University of North Carolina for critically evaluating the manuscript.

References

Josefsson LG and **Rask L** (1997) Cloning of a putative G-protein-coupled receptor from *Arabidopsis* thaliana Eur. J. Biochem **249**(2): 415-420.

Krysan PJ, Young JC and **Sussman MR** (1999) T-DNA as an Insertional Mutagen in *Arabidopsis*. Plant Cell **11**: 2283-2290.

Ma H, Yanofsky MF and **Meyerowitz EM** (1990) Molecular cloning and characterization of GPA1, a G protein alpha subunit gene from *Arabidopsis* thaliana. Proc. Natl. Acad. Sci. USA **87**(10):3821-

Weiss CA, Garnaat CW, Mukai K, Hu Y and **Ma H** (1994) Isolation of cDNAs encoding guanine nucleotide-binding protein beta-subunit homologues from maize (ZGB1) and *Arabidopsis* (AGB1). Proc. Natl. Acad. Sci. USA, Sept. 27, 1994; **91**(20): 9554-8

Reference on Internet resources

Baxevanis AD (2001) The molecular biology database collection: an updated compilation of biological database resources. Nucleic Acids Research **29**(1):1-10.

Kaminski N (2000) Bioinformatics. Am. J. Respir. Cell Mol. Biol. **23**: 705-711.

Peruski LF and **Peruski AH** (1998) The Internet and the New Biology, Tools for Genomic and Molecular Research. American Society for Microbiology (Book)

Rhee, SY. 2000. Bioinformatics resources, Challenges, and Opportunities using *Arabidopsis* as a model organism in a post-genomic era. Plant. Physiol. **124** : 1460-1464.

Somerville C, Flanders D and **Cherry JM** (1997) Plant Biology in the Post-Gutenberg Era. Plant Physiol. **113** : 1015-1022.