U-1
Electives

Genes & Genome Studies along with Applications in ~~Biomedical Studies~~. Biomolecular Diseases.

~~String Database~~

- Genes – basic and functional unit of heredity.
  - made up of DNA.
  - types – A, T/U, G, C.
  - Seq. within NA → represents a single protein
  - Gene may exist in alternative forms → alleles.

* INTRONS –

- (intragenic region) – any nucleotide seq. within a gene that is removed by ~~RNA~~ splicing during maturation of the ~~mRNA~~ final RNA product.
- Non-coding regions of an RNA Transcript, or the DNA encoding it, that are eliminated by splicing before translation.
- Functions – Transcription termination, Genome organization, Transcription initiation etc.
- Adv. – Evolutionary adv. of introns include – the possibility to create new genes by cutting and pasting exons from existing genes or to diversify the protein O/P of a single gene by splicing the exons together in diff. ways.
- What happens if introns are not removed?
  - Introns carry information to build a protein + they actually have to be removed in order for the mRNA to encode a protein with the right seq.

PAGE No.
DATE : / /

* Exons -
- any part of a gene that will encode a part of the final mature RNA produced by that gene after introns have been removed by RNA splicing.
- refers both the DNA seq. within a gene and to the corresponding seq. in RNA transcripts.
- How many exons does a gene have? — 8.8 exons
  - On an avg. there are 8.8 exons and 7.8 introns per gene.
  - About 80% of the exons in each chromosome are < 200

* TOOLS —
(1) Genomescan (2) Genezid (3) GrailEXP

* ORF [Open Reading Frame] — (#) All ORF can be exons. All exons cannot be ORF.
- portion of a DNA molecule that, when translated into a.a., contains no stop codons.
- A long ORF is likely part of a RF that has the ability to be translated.
- An ORF is a continuous stretch of codons that may begin with a start codon (usually AUG) and ends at a stop codon (usually UAA, UAG or UGA).

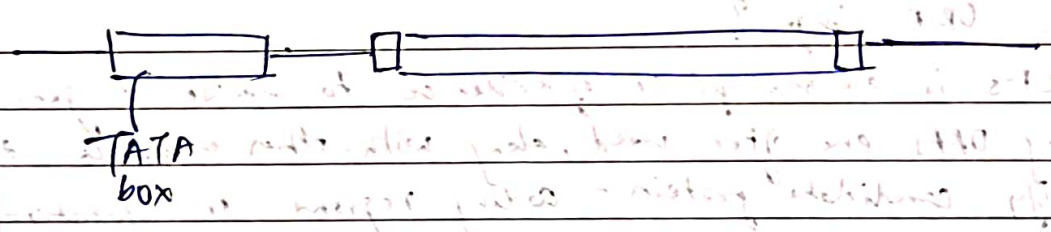
- Why are ORF imp?
  ↳ ORFs is as one piece of evidence to assist in gene prediction.
  ↳ long ORFs are often used, along with other evidence, to initially identify candidate protein-coding regions or functional RNA-coding regions in a DNA sequence.

Tools - NCBI - ORF FINDER.

ELECTIVES CONTINUED

– Identifying ORFs –

- Simple 1st step in gene findings
- Translate genomic seq. in six frames.
- Identify stop codon in each frame.
- Regions without stop codons are called "open reading frame" or ORFs.
- Locate and tag all of the likely ORFs in a sequence.
- The longest ORF from a methionine codon is a good prediction of a protein encoding sequence.



PAGE No.
DATE / /

* Promoters -
- The DNA opens up in the promoter region so that RNA polymerase can begin transcription.
- Each gene (or in bacteria, each grp of genes transcribed together) has its own promoter.
- A promoter contains DNA seqs. that let RNA polymerase or its helper proteins attach to the DNA i.e., A DNA seq. that the transcription apparatus recognizes and binds.
- Usually found near the beginning of a gene, the promoter has a binding site for the enzyme used to make a mRNA molecule.

(1) TTS - Transcription Terminator Site
TSS - —11— Start —11—

(2) Diagram - Transcription unit

* Tools (Eukaryotic + Prokaryotic Tools) -
(1) FPROM (2) TSSP (3) TSSW (4) TSSG (5) BpROM

* Splice sites -
- A genetic alteration in the DNA seq. that occurs at the boundary of an exon and an intron (splice site).
- Splice sites are the seqs. immediately surrounding the exon intron boundaries.
- Found - These sites are found at the 5' and 3' ends of introns. Most commonly, the RNA seq. that is removed begins with the dinucleotide GU at its 5' end, and ends with AG at its 3' end.
i.e., the GU-AG rule (originally called the GT-AG rule in terms of DNA sequence).

PAGE No.
DATE / /

- Why splice sites are imp?
  - Mutations in these seqs may lead to retention of large segments of intronic DNA by the mRNA, or to entire exons being spliced out of the mRNA. These changes could result in prod^n of a non-functional protein.
  - These donor sites, or recognition sites, are essential in the processing of mRNA.

Diagram Normal pre-mRNA

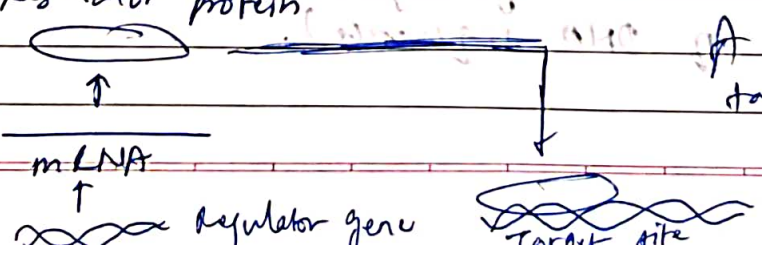
- Why splice sites prediction are imp?
  - Prediction of splice sites where accurate localization of splice sites can substantially help explore the str. of genes.
  - Accurate prediction of splice sites can setup the boundaries of exons which is critical in alternative splicing prediction.

- Tools (Eukaryotic + Prokaryotic Tools) (Algo based on various Approach / Algorithm) —
  (1) Human Splicing Finder (2) GeneSplicer
  (3) FGENES (4) Fgenesh-M (5) FGENESH-GC

Regulatory Regions —

- A segment of a NA molecule which is capable of increasing or decreasing the expression of specific genes within an organism.
- An enhancer activates the nearest promoter to it.
- A UAS [upstream activator seq.) in yeast behaves like an enhancer but works only upstream of the promoter.
- Form complexes of activators that interact directly or with the promoter.

Regulatory model



A regulator binds a target site on DNA.

// ① the location of ORFs
② Predict the structures of introns as well as exons if the genes of interest are of eukaryotic region.

- What is the role of a regulatory region?
- Regulatory seq. controls when expression occurs for the multiple protein coding regions (red)
- Promoter, operator and enhancer regions (yellow) regulate the transcription of the gene into an mRNA.
- The mRNA untranslated regions (blue) regulate translation into the final protein products.

- Tools (Euk + Prokaryotic Tools) (Also based on various approach / algorithm) —
① TRANSFAC ② RSAT Fungi / prokaryotes / bacteria
③ CREME ④ RSA Tools

* Prediction for various signals in Genome is important?

- Why gene prediction?
With the rapid accumulation of genomic seq info, there is a pressing need to use computational approaches to accurately predict gene structure.
- Computational gene prediction is a pre-requisite for detailed functional annotation of genes and genomes.
- The process includes detection of— //

- * The Ultimate goal —
- To describe all the genes computationally with near 100% accuracy.
- The ability to accurately predict genes can significantly reduce the amount of experimental verification work required.

- Disadvantage —
- For eukaryotes, ∴ many problems in computational gene prediction are still largely unsolved.
- This is because coding regions normally do not have conserved motifs.
- The elements are diverse and not clearly defined.
- Detecting coding potential of a genomic region has to rely on finding features associated with genes that may be very difficult to detect.

PAGE No.
DATE / /

* • Normally elements are short (6-8 nucleotide) and found in any seq. by random chance thus ↑ high rate of false (+ve) results because of which sensitivity drops and specificity is hampered.

- Soln -

For preliminarily identification of these elements → combine a multitude of features and use sophisticated algorithms that give either —

(1) ab initio - based predictions or
(2) Predictions based on evolutionary info. (Homology based) or
(3) Experimental data.

* Types of Approaches —

(1) Ab initio - based (prediction based on given seq. only) :- Relying on 2 major features associated with genes —

First feature — is the existence of gene signals, which include start and stop codons, intron splice signals, transcription factor binding sites, ribosomal binding sites and polyadenylation (poly-A) sites, the triplet codon etc.

— Thus, unique features can be detected by applying probabilistic models such as Markov models or Hidden Markov models, MLT to help distinguish coding from non-coding regions.

(2) Homology - based approaches

- Predictions based on significant matches of the query sequence with sequences of known genes.

eg: if a translated DNA seq is found to be similar to a known

Therefore, There are also a no. of programs that actually combine prediction results from multiple individual programs to derive a consensus prediction. This type of algorithms can therefore be considered as consensus - based.

* Gene Prediction in Prokaryotes -

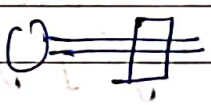
- Prokaryotes, which include bacteria and Archae, have relatively small genomes with sizes ranging from 0.5 to 10 Mbp ($1 Mbp = 10^6 bp$).
- The gene density in the genomes is high, with more than 90% of a genome sequence containing coding seq.
- There are very few repetitive sequences.

- Occasionally GTG and TTG are used as alternative start codons, but methionine is still the actual a.a. inserted at the first position.

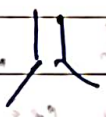
- There may be multiple ATG, GTG or TGT codons in a frame.
- But presence of these codons at the beginning of the frame does not necessarily give a clear indication of the translation initiation site.
- Instead, to help identify this initiation codon, other features associated with translation are used.

- In many bacteria, it has a consensus motif of AGGAGGT.
- Identification of the ribosome binding site can help locate the start codon.

Transcription start



RBS



- At the end of the protein coding region is a stop codon that causes translation to stop.
- Many prokaryotic genes are transcribed together as 1 operon
- The end of the operon is characterized by a transcription termination signal → ρ - Independent terminator.

PAGE No.
DATE / /

* Conventional Determination Method of ORF —

- Without use of specialized programs, prokaryotic gene identification can rely on manual determination of ORFs and major signals related to prokaryotic genes.
- Prokaryotic DNA is first subjected to conceptual translation in all 6 possible frames, three frames forward and three frames reverse.
- Because a stop codon occurs in about every 20 codons by chance in a non-coding region, a frame longer than 30 codons without interruption by stop codons is suggestive of a gene coding region.

N/E

- In the early stages of development of gene prediction algorithms, genes were predicted by examining the non-randomness of nucleotide distribution.
- One method is based on the nucleotide composition of the 3rd position of a codon. Thus, in coding seq, it has been observed that this position has a preference to use G or C over A or T. (Ⓐ Klobkle hypothesis).

N/E

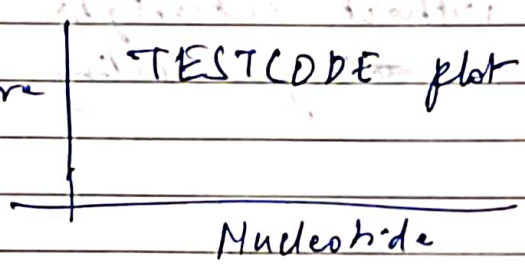
GC content plot

Relative probability

GC bias

- In addition to codon bias, there is a similar method called TESTCODE that exploits the fact that the of the third codon nucleotides in a coding region tend to repeat themselves.
- By plotting the repeating patterns of the nucleotides at this position, coding and non coding regions can be differentiated.

NK



- Thus, statistical methods, which are based on empirical rules, examine the statistics of a single nucleotide (either G or C).
- They identify only typical genes and tend to miss atypical genes in which the rule of codon bias is not strictly followed.

NK

* Gene Prediction Using Markov Models and Hidden Markov Models —

- MM and HMM can be v. helpful in providing fines statistical description of gene.
- MM - describes the probability of the distribution of nucleotides in a DNA seq, in which the conditional probability of a particular seq. position depends on k previous positions.

- In this case, k is the order of a MM.

NK

- A second-order model looks at the preceding 2 bases to determine which base follows, which is more characteristic of codons in a coding seq.

- The use of MM in gene finding exploits the fact that oligonucleotide distribution in the coding regions are diff. from those for the non-coding regions.
- these can be represented with various orders of MM.
- Since a fixed-order Markov chain describes the probability of a particular nucleotide that depends on previous k nucleotides, the longer the oligomer unit, the more non-randomness can be described for the coding region.

NK

- Once the parameters of the model are established, it can be used to compute the non-random distributions of primers or hexamers in a new seq. to find the regions that are compatible with the statistical profiles in the learning set.

Gene prediction — Xin Xiong

PAGE No.
DATE / /

- The frequency of 6 unique nucleotides appearing together in a coding region is much higher than by random chance.
- ∴ a fifth-order MM, which calculates the probability of hexamer bases, can detect nucleotide correlations found in coding regions more accurately.
- Problem of fifth-order Markov chain = if not if there are not enough hexamers which happens in short gene sequences, the method's efficacy may be limited.

NIL

- Sometimes, genes tend to escape detection using the typical gene model. Thus, to make the algorithm capable of fully describing all genes in a genome, more than one MM is needed and therefore HMM prediction algorithms are implemented

* Performance Evaluation —
Accuracy of a prediction program — Sensitivity & Specificity
— calculated on 4 features accurately which are —

(1) TP — correctly predicted feature
(2) FP — incorrectly " "
(3) FN — missed feature
(4) TN — correctly predicted ∄ absence of a feature.

PAGE No
DATE / /

(1) Specificity (Sp) = TP / (TP + FP) (Proportion of true signals among all signals that are predicted i.e., an ability to exclude incorrect predictions).

- A program is considered accurate if both sensitivity and specificity are simultaneously high and approach a value of 1.
- If the sensitivity is high but specificity is low, the program is said to have a tendency to over predict.

NK

- In the field of gene finding, a single parameter known as the correlation coefficient (CC) is often used, which is defined by

$$CC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TN + FN)(FP + TN)}}$$

- CC provides an overall measure of accuracy, which ranges from -1 to +1, with
  +1 = always correct prediction
  -1 = " " incorrect " "

* Gene Prediction in Eukaryotes -

NK

PAGE No.
DATE / /

- The space b/w genes is often very large and rich in repetitive seqs and transposable elements.
- Eukaryotic genomes are characterized by a mosaic organization in which a gene is split into pieces (exons) by intervening non-coding seqs (introns).
- A eukaryotic gene is modified in 3 diff. ways before becoming a mature mRNA for protein translations.
  a. Capping at the 5' end of the transcript (methylation of the initial residue of the RNA).

NC

- Computationally very demanding because of the presence of split gene structures, alternative splicing and very low gene densities.

- Solⁿ - some conserved seq. features in eukaryotic genes that allow the computational prediction.

Eg:- the splice junctions of introns and exons follow the GT-AG rule in which an intron at the 5' splice junction has a consensus motif of GTAAGT; and at the 3' splice junction is a consensus motif of (Py)12 NCAG.

NK

- In addition, most of these genes have a high density of CG dinucleotides near the transcription start site.
- And this region is referred to as a CpG island (p refers to the Phosphodiester bond connecting the 2 nucleotides), which helps to identify the transcription initiation site of a eukaryotic gene.

- The poly-A signal can also help locate the final coding sequence.

* Gene Prediction Programs –

- 3 categories of algorithms
  1. Ab initio based
  2. Homology based
  3. Consensus based

(1)
a. Prediction using Neural Network
b. Using Discriminant Analysis
c. Using HMM's.

(1)

a. Gene signals – includes signals such as gene start & stop sites and putative splice sites, recognizable consensus seqs such as poly-A sites.

b. Gene content – includes coding statistics, such as non-random nucleotide distribution, a.a. distribution, synonymous codon usage and hexamer frequencies

- Neural Network / ANN for gene prediction –

↳ Another feature that it resembles to nervous system is – its ability to "learn" and then make predictions after being trained.

↳ The network is able to process info. and modify parameters of the weight functions b/w variables during the training stage. Once it is trained, it is able to make automatic predictions about the unknown.

PAGE No.
DATE / /

↳ A NN is constructed with 3 layers -

//

NK

↳ The gene str. info. is separated into several classes of features such as hexamer frequencies, splice sites and GC composition during training.

↳ The weight functions in the hidden layers are adjusted during this process to recognize the nucleotide patterns and their relationship with known structures.

NK

- Tool - GRAIL

(i) b.

~~that best separation~~

// and ∴ any identified variation could be responsible for problems.

PAGE No.
DATE / /

Genome Analysis, Assembly & Annotation Studies

* Synteny -

(NK)

Use - provides a framework in which conservation of homologous genes and gene order is identified b/w genes of diff. species.
- The availability of human and mouse genomes paved the way for algorithm development in large scale called as, synteny mapping, which eventually became an integral part of comparative genomics.

* What is synteny mapping?

(NK)

- Gene order conservation is in fact rarely observed among divergent species.
- ∴, comparison of syntenic relationships is normally carried out b/w relatively close lineages.
- However, if syntenic relationships for certain genes are indeed observed among divergent prokaryotes, they often provide imp. clues for functional relationships of the genes of interest.

Eg - Genes involved in the same metabolic pathway tend to be clustered among phylogenetically diverse organisms. The preservation of the gene order is a result of the selective pressure to allow the genes to be co-regulated and function as an operon.

* Reference genome seq. / Reference genome assembly —

(NK)

Eg - The human reference genome doesn't represent the genetic seq. for any one individual, but is made up of a combination of several people's DNA. When we sequence a patient's genome and compare it to the reference genome, we assume that the reference represents the 'normal' sequence ...

Contd..→

Expensive technique (disadv).

PAGE No. / DATE / /

How to search = (N/K)

① * Gene expression profiling - (N/K)

- Methods for analysis -
  (1) Microarray # (2) RT-PCR (3) PAGE
  (4) Northern blot (5) ELISA Test (6) Western Blot etc.

→ Advantage -
(1) Understand molecular mechanisms of complex disorders.
(2) Cancer studies
(3) Metabolic disorder studies etc,

* Structural variants of DNA

(1) Relevance in molecular level processes
(2) Identification
(3) Assembly of data from genome sequencing

- Systems and gene order - Any mol. bio book
- Mutation and genetic disease - refers to chromosomal studies
  → Case study approach. (Mendelian Inheritance problem)
- Integrated genomic maps - ~~proper~~ (for OMIM database
  Genomic + Proteomics
  or
  only Genomics
  or
  only Proteomics
  Bioinformatics, Methodology)
  O/P of that research.

PAGE No.
DATE / /

* Structural Variants –

(NK)

- Microscopic & submicroscopic structural variant such as deletions, duplications, large copy no. variants as well as insertions, inversions and translocations.

- * What is Variant ~~data~~ DNA –

(NK)

* • Variant can be used to describe an alteration that may be pathogenic / of unknown significance.
• May be germline or somatic.

- Why are structural Variants imp?

(NK)

↳ Regulation of gene expression, ethnic diversity and large-scale chromosome evolution – giving rise to the differences within populations and among

- How do you identify structural variations –

(NK)

• It is widely agreed that the method into 4 algorithms – can be classified

Read-pair (RP), Split-Read (SR), Read-Depth (RD) and Assembly

- What seq. tech. can be used to detect SV?

(NK)

• Strand-Seq is the most suitable detection method for chromosomal inversions, a particularly challenging type of SV.

PAGE No.
DATE / /

* DNA sequencing –

(1) Maxam-Gilbert : – >

(2) Sanger – key reagent – ddNTP . – –

(3) Restriction endonucleases – Bacterial enzymes that cleave DNA at specific sites.

* NGS –
(NK)

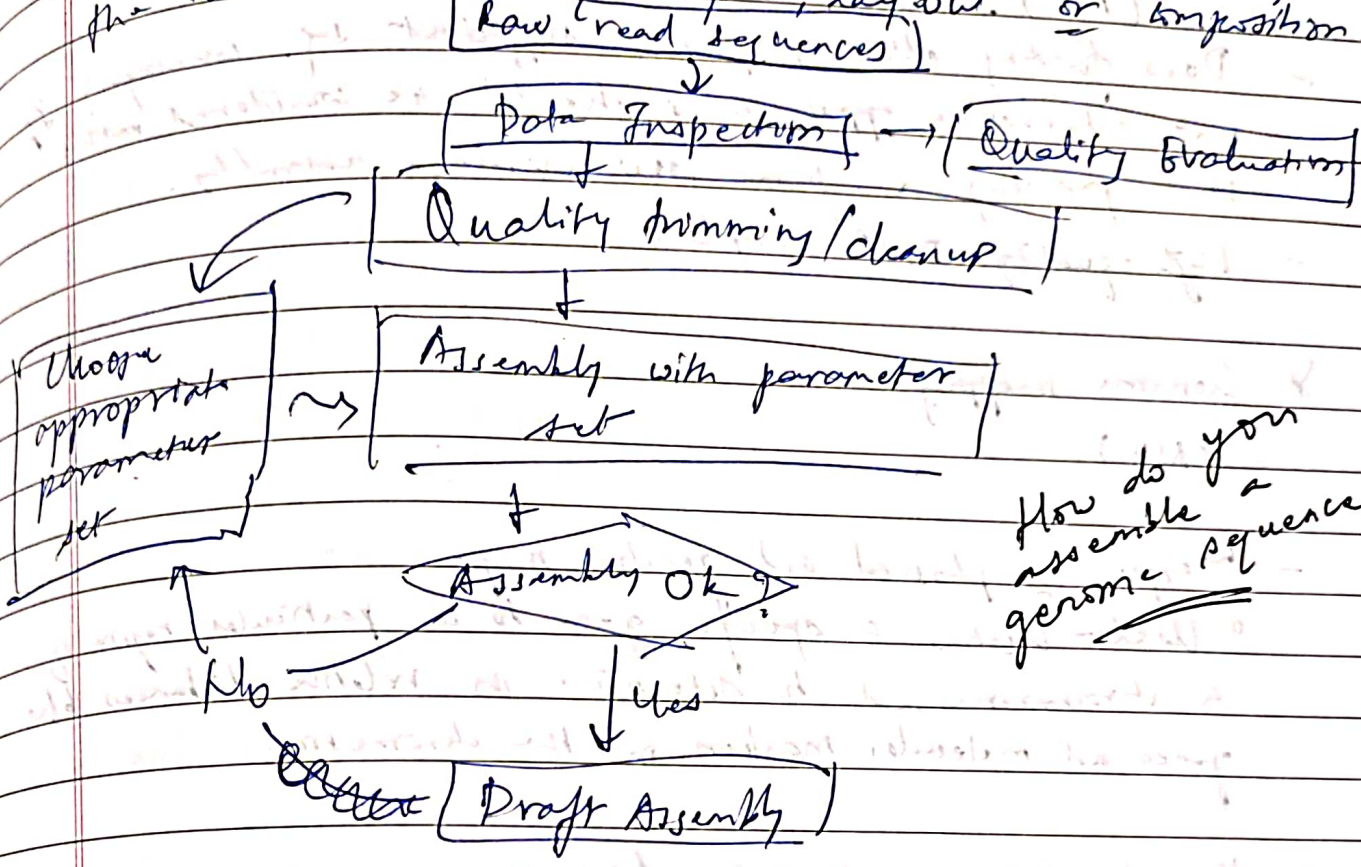
- NGS is a powerful platform that has enabled the sequencing of 1000s to million of DNA molecules simultaneously.
- This powerful tool is revolutionizing fields such as – personalized medicine, genetic diseases and clinical diagnostics by offering a high throughput option with the capability to sequence multiple individuals at the same time.
- Have gained an understanding of the underlying cause of disease.

* Assembly of Data from Genome Sequencing –
(NK)

- In Bioinformatics, seq. assembly refers to aligning and merging fragments from a longer DNA seq. in order to reconstruct the original seq. (typically the short fragments called read, result from shotgun sequencing Genomic DNA or gene transcript (EST).

PAGE No.
DATE / /

- De Novo Gene Assemblies assume no prior knowledge of the source DNA eg. length, layout or composition.



How do you assemble a genome sequence?

Steps - (NK - ①, ②, ③)

① Choose the best sequencing platforms and library preparations.
② Select the best possible DNA source and DNA extraction method.
③ Check the computational resources and requirements.

NK - ④, ⑤, ⑥ ⑦ Assemble the genome.
⑧ Check the assembly quality before annotation.
⑨ Genome annotations
⑩ Build a searchable and shareable Output format.
⑪ Reach out to the community to refine the assembly and annotation.

* Conclusions -

(NK)

Beginners & small resea

PAGE No.
DATE / /

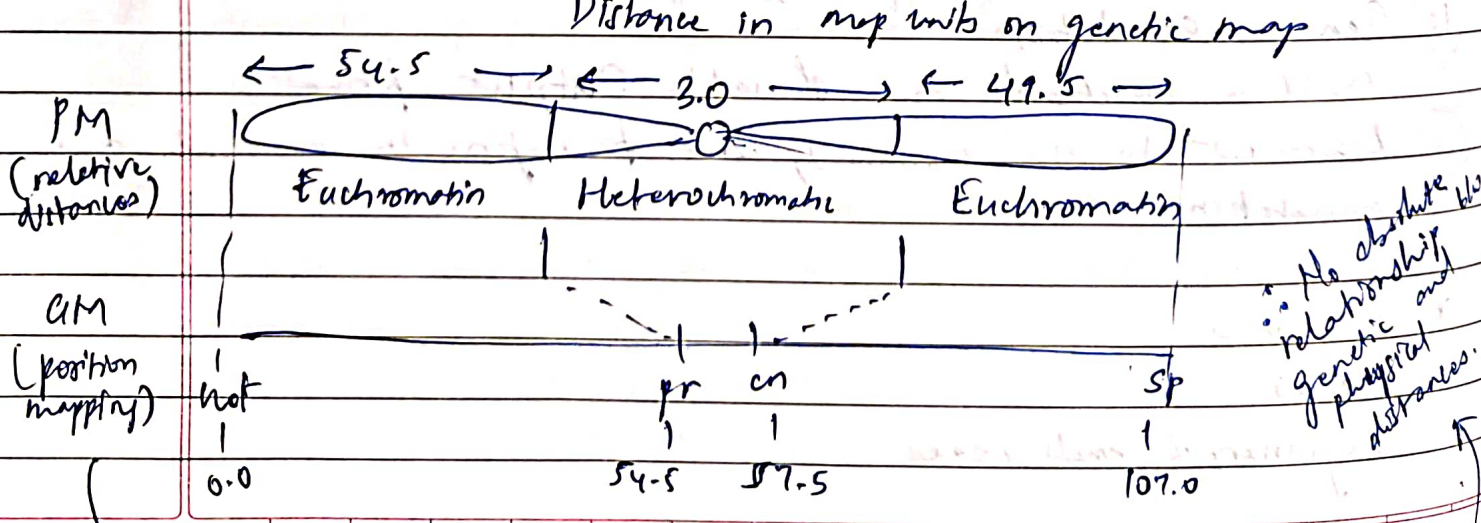
* Imp. points to consider -
(NK).
- Does funding allow to produce sufficient seq. coverage? If not, alternative approaches should be considered rather than producing a poor low coverage assembly.
- High-quality DNA sample.

* Genome mapping -
(NK)

- Mapping of physical and genetic maps
- Used - locating a specific gene to a particular region of a chromosome and to determine its relative distances b/w genes and molecular markers on the chromosome.

- How is genomic mapping done -
(NK)
To do this, the genome is first broken up into fragments. The fragments are then replicated upto 10 times in bacterial cells to create a library of DNA clones.

- Genetic vs Physical Distance - Diagram



data may have limited resolution & limited accuracy (recombinational "hotspots" or "coldspots" (not enough problem)

PAGE No.
DATE / /

* Comparing GM and PM -
- Saccharomyces cerevisiae chromosome III.
- physical map obtained by DNA sequencing.
- the order of the upper 2 markers (glk1 and cha1) is incorrect on the GM.

- Approach for Genome Analysis -
ESTs, shotgun, DNA markers (RFLP, SSLP, SNP), PCR, Microarray (Protein MA, DNA MA), Restriction mapping, FISH, STS, EST.

PMID - 17703239
doi - 10.1038/nrg2144

* Identification of Disease Genes & Role of Bioinfo - eg - Database - OMIM

4FO → microarray maps
Green - regular expression genes (no mutations)
Red - mutated &.
Black
Yellow

Gene Prediction - Continued FP = false positive.

* Prediction of Promoter and Regulatory Elements - (PRE) (PRE)

:. These DNA elements directly regulate gene expression.
- PRE are traditionally determined by experimental analysis, but time-consuming and laborious.
-:. Computational approach to solve the problem and enhance the potential. -

- Problems - (PRE)

Reasons -

(3) Lack of data.
(4) PRE cannot be translated into protein sequences to ↑ the sensitivity for their detection.
(5) Normally elements are short (6-8 nucleotide) and found in any seq. by random chance :. ↑rate of FP results because of which sensitivity drops and specificity is hampered.

- Soln -

(4) Predictions based on evolutionary info.
(5) Experimental data.

PAGE No.
DATE / /

* PRE in prokaryotes -

(1) Transcription - RNA polymerase
σ 70 → recognizes specific sequences upstream of a gene and allows the rest of the enzyme complex to bind.
-35 and -10 boxes
(└ TATAAT
consensus seq - TTGACA
(AAK)

(2) Operon
- Transcription factors (TF) - bind to specific DNA seqs. do either enhance or inhibit the function of the RNA polymerase.
- Regulatory elements. - specific DNA seqs to which TF bind.
↳ may bind in the vicinity of the promoter or bind to a site several 100 bases away from the promoter.
- Regulatory proteins binding @ long distance can still exert their effect is because of the flexible structure of DNA which is able to bend and exert its effect by bringing the transcription factors in close contact with the RNA polymerase complex.

* PRE in Eukaryotes -

(1) RNA polymerase I - transcription of rRNAs and tRNAs respectively
(2) II - transcribing protein - encoding genes for synthesis of mRNAs)
(3) III - [illegible] (same as I)

- Each of them transcribes diff. sets of genes.
- Unlike prokaryotes.

PAGE No.
DATE / /

RNA pol II does not directly bind to the promoter, but relies on a dozen or more transcription factors to recognize and bind to the promoter in a specific order before itself.

(NK)

- Many genes have a unique initiator sequence (Inr) which is a pyrimidine rich seq with a consensus (C/T)(C/T) CA (C/T)(C/T)
- This site coincides with the transcription start site.
- Mostly TF binding sites are located within 500 bp upstream of the Transcription start site.
- Some regulatory elements sites can be found 10 of 1000 bp away from the gene start site.
- RE are located downstream instead of upstream of the transcription start site.

* Prediction Algorithms — (NK)

(3) Expression profile based — Many profiles constructed from a no. of coexpressed gene seqs from the same organism.

(4) Phylogenetic footprinting.

- RNA pol II describes the eukaryotic mRNA

(1) Ab-Initio based — (NK)

- Conventional Method — Detection of element via matching a consensus seq. pattern represented by regular expression or matching a ⟶ position - specific scoring matrix (PSSM).

- But in other case, the consensus seqs of the matrices are relatively short, covering 6–10 bases and generates ↑ rate of FP as a result.
- Sol^n – New algorithm based on NN, ANN, HMM, MLT. – –

(a) Prediction for Prokaryotes → (NN)

- But set of simple rules (Mahony et al) is accurate which relies on < kinds of info —
  - gene orientation and intergenic distances of a pair of genes of interest and conserved linkage of the genes based on comparative genomic analysis i.e., gene linkages are focused.
- A scoring scheme is developed to assign operons with diff. levels of confidence.

CLAIMS – Accurate identification of an operon str., which in turn facilitates the promoter prediction.
- This authority scheme not yet available as a computer program thus can be done manually using the rules, but few programs doesn't accept this approach.
- Scoring criteria for operon prediction
- Prediction Tool for prokaryotes – BPROM.

Prediction for Eukaryotes -
(Alk).

- Also because of the ↑ variability of TF Binding sites, the simple seq. matching often misses true promoter sites, creating FN.
- ∴ to ↑ specificity in prediction unique feature of eukaryotic promoter is exploited, which is the presence of CpG islands (many vertebrate genes are characterized by a ↑ density of CG dinucleotides near the promoter region overlapping the TSS)
- Eg ∴ CpG Island identification can lead to true promoter in the immediate upstream

Prediction Tool - TSSW

(c) Phylogenetic Footprinting - Based Method -
(Alk)

- Caveats / Red flags are -
(1) Organisms selected are too closely related (human & chimpanzee) the seq. diff. bw them may not be sufficient to filter out functional elements.
(2) Organisms evolutionary distances are too long (such as human & fish), long evolutionary divergence may render promoter and other elements undetectable.
(3) To extract non-coding seqs upstream of corresponding genes and focus the comparision to this region only, which helps to prevent FPs.
(4) Predictive value depends on the quality of the pairwise seq. alignment ∴ Advance alignment program is used.

PAGE No.
DATE / /

- ∴ Software programs are specifically designed to take advantage of the presence of phylogenetic patterns to make comparisons among a no. of related species to identify putative & TFBS.
- Adv – No having of probabilistic models req^d.

(3) Expression Profile Based Method –

Xin Xiong

DNA Microarray Introduction

(4) Affimitrix chips – Xin Xiong 26

* Basic Concepts –

- Prokaryotic v/s Eukaryotic cells
- Cell Organelles ; Chromosomes
- DNA – genetic maps
- RNA
- Central Dogma

- The Goal –
(NK)

One approach – What happens in Time

Large scale approaches – Now (NK)
↳ Microarrays allow massive parallel measurements in one experiment

- The Southern Blot –
- Microarray – (NK) –
↳ Array – Orderly arrangement of large gene samples.

PAGE No.
DATE / /

- Background -
  - Functional genomics - (NK)
    - Study of obtaining an overall picture of genome functions, including the expression profiles at the mRNA level and the protein level. (NK)
    - Integrative biology & systems biology studies to understand health & diseased states.
    - Drug discovery process.
  - Gene expression - NK
  - Microarray -
    - Tool is used to measure the presence & ~~dosage~~ abundance of gene expression in tissues.

- Outline of Microarray - (NK)

(I) Interpretation of result
  a. Statistical analysis --

* What is Microarray - (NK)

- In particular, the amount of mRNA for each gene in a given sample (or a pair of samples) is measured -
  - Parallel
  - High - Throughput
  - Large scale
  - Genomic state

- Historical perspective (NK)
  - Fixing samples

PAGE No.
DATE / /

(NK) process of copying of DNA into mRNA
- Transcription - Environment dependent.
- Microarray detects mRNA, or rather the more stable cDNA.

* Central Dogma -

* Complement Hybridization -
- Due to Watson-Crick base pairing, an mRNA molecule will hybridize to a complementary DNA molecule.

* Procedure of Microarray -
(NK)
(3) Scanning → Scan
(4) Analysis

* Criteria for Designing
* Software to be used - Oligo Wiz, Oligo Array

* Microarray or DNA chip.
* Data Collection
* Visualizing bound cDNA.
* Analyzing the data
* Image processing.

* Data Transformation + Normalization → involves mathematical calculations - Cy3 v/s Cy5 graph → Normalized using log 2; v/s fold; no fold statistics

* Sample Preparation
* Hybridization
* Scanning

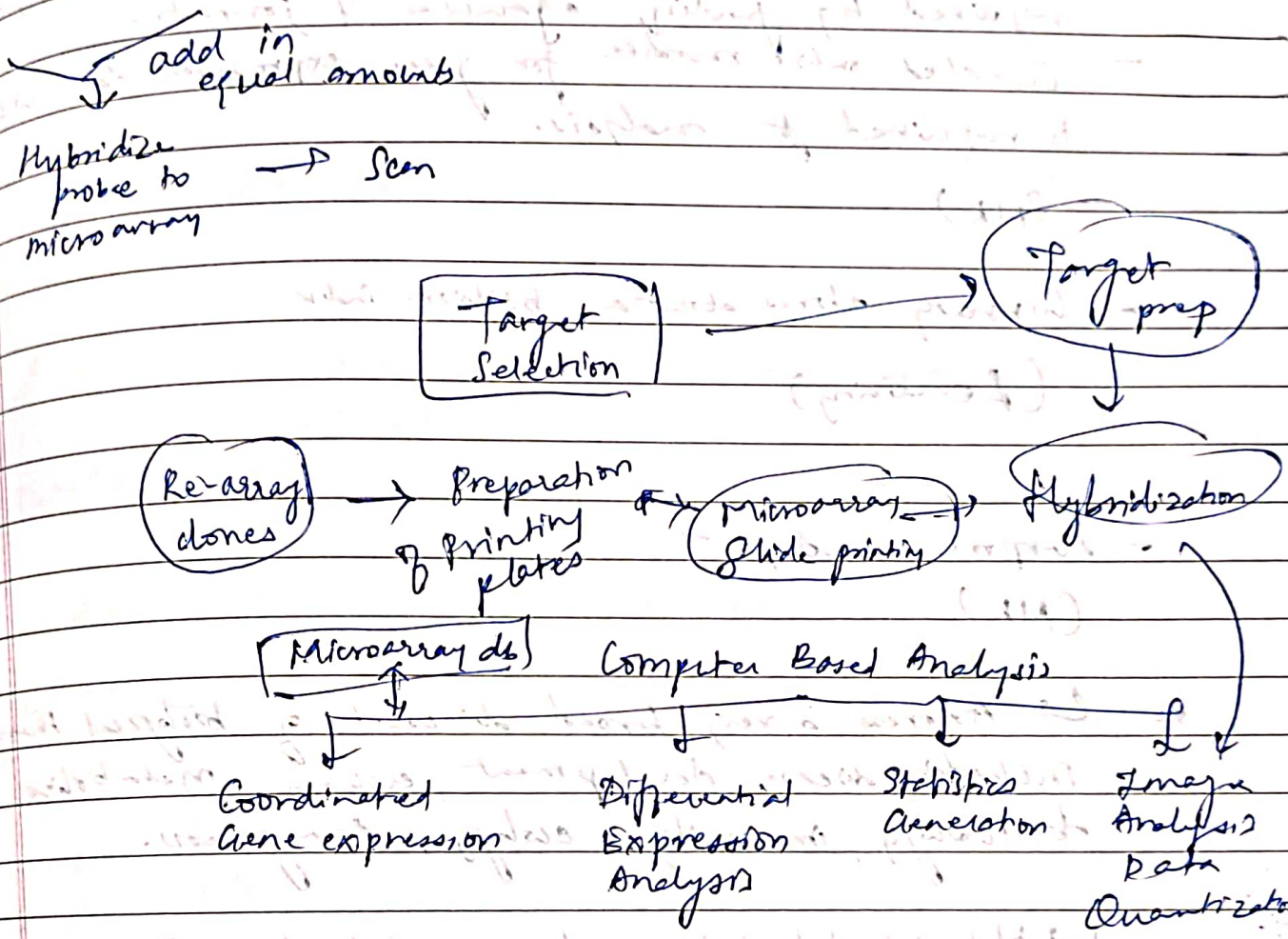
* What is Clustering ? → Partitioned Algorithms; Hierarchical Algorithms.
- Partitioned: Non-hierarchical, non-overlapping clusters
- Hierarchical: Bottom up (agglomerative); Top-Down (divisive)

Standard → green color
diseased → red

Applications

GEO [Gene Expression Omnibus]

Also called "Microarray Databases"



* Output of Microarray -

(NK)

- No immediate observations can be made concerning gene expression from raw data.
- Statistical analysis applications are used to interrogate the data for information on gene expression patterns.

PAGE No.
DATE / /

* GEO

(MK)

- Convenient for deposition of gene expression data, as required by funding agencies & journals.
- Curated, online resource for gene expression data whose is required for analysis.

(MK)

- Currently stores about a billion indiv.

(Remaining)

- Purpose and Scope -

(MK)

↳ Address a very broad diversity of biological themes including disease, development, evolution, metabolism, toxicology, immunity, ecology, transgenesis.

↳ Not intended to be used as a Laboratory Information Management System (LIMS) or a pre/first analysis

- 3 main goals of GEO -

(1) MK

(2) Offer simple submission procedures and formats that support complete and well-annotated data deposits from the research community (submission guide).

(3) Provide user-friendly mechanisms that allow users to query, locate, review and download studies

PAGE No.
DATE / /

* GEO structure / Architecture -

1.
2.
3.

- 4 kinds of data records.
(NK)

(3) Series (GSE) = defines a set of samples and how they are related.

(4) DataSets (GDS) = sample data collections assembled by GEO staff.

- ↳ Submitters may provide raw data.

Submitted by manufacturer	Submitted by Experimentalists		Curated by NCBI
GPL	GSM	GSE	GDS
Platform description	Raw / processed spot intensities from a single slide / chip	Grouping of slide / chip data in a single experiment	Grouping of experiments

Data Organization

Platform → Sample → Series

↓

Dataset → Profile

PAGE No.
DATE / /

* Navigatory GEO -

G-Protein Coupled Receptor

(NIK)

- Belongs to a family of protein that act like a molecular switches inside the cells and are involved in transmitting the signal from a variety of stimuli outside a cell to the inside.

-

(NIL)

- When an external signaling molecule binds to a GPCR, it causes a conformational change in the GPCR.
- This change then triggers the interaction b/w the GPCR and a nearby protein.

* Structure -

(NIK)

* Types of GPCR -

(1) Class A – Rhodopsin-like
B – Secretin-like
C – Metabotropic glutamate/pheromone
D – Fungal pheromone
E – cAMP receptors

main classes based on sequence homology and function.

(2) Frizzled / Smoothened family

(3) Putative families
↳ Olfactory odorant proteins
↳ Insect odorant receptors
↳ Plant Mlo receptors
↳ Nematode chemoreceptors
↳ Vomeronasal receptors (V1R and V3R)
↳ Taste receptors (TAR)

(4) Orphans
↳ Putative/ undesignated GPCRs

(5) Non-GPCR families
↳ Archael / Bacterial / fungal opsins

PAGE No.
DATE / /

* Types of ligands

(1) Agonists
- ligands which shift the equilibrium in favor of the active state.

(2) Inverse Agonists
- Ligands which shift the equilibrium in favor of inactive states

(3) Neutral Antagonists
- ligands which do not affect the equilibrium.

* Role of GPCR - (in body)
(*) Hormones (*) Odorants (*) Tastes

* GPCR function -

(1) Visual sense - Opsins use a photoisomerization rxn to translate electromagnetic radiation into cellular signals. Rhodopsin, for eg, uses the conversion of 11-cis-retinal to all-trans-retinal for this purpose.

(2) Sense of smell - Receptors of the olfactory epithelium bind odorants (olfactory receptors) and pheromones (vomeronasal receptors).

(3) Behavioral & Mood regulation - Receptors in the mammalian brain bind several diff. neurotransmitters, including serotonin and dopamine.

(4) Regulation of immune system activity and inflammation -
Chemokine receptors bind ligands that mediate intercellular communication b/w the immune system; receptors such as histamine receptors bind inflammatory mediators & engage target cell types in the inflammatory response.

(5) Autonomic Nervous System - Both the sympathetic and parasympathetic nervous systems are regulated by GPCRs pathways. These systems are responsible for control of many automatic functions body such as blood pressure, heart rate, digestive -

PAGE No.
DATE / /

* oGPCR [Orphan GPCR] -

- GPCR belongs to a superfamily and ∴ would share seq. similarities.
-
- The no. of GPCR is about 800, of which more than half are olfactory GPCRs.
- The discovery of new GPCRs found by homology screening suffers from the obvious problem, the receptors found lack their pharmacological identities, their natural ligands.
- Without knowing what their endogenous ligands are, we lack the information to understand their physiological role and hence cannot...
- Traditionally the ligand were found first and served to characterize the receptors pharmacologically.
- But by the end of the 1980s, study for identification of GPCRs was started by using homology screening approaches.

BUT

- GPCRs found many homology approaches lack their natural ligands, thus called as "Orphan" GPCRs.
- Search for ligands of orphan GPCRs has given birth to the "REVERSE PHARMACOLOGY" approach.
- ∴ oGPCR as targets to identify their endogenous ligands started.

Orphan receptor

↓

Similar structure to other identified receptors but whose endogenous ligands has not yet been identified.

↓

If a ligand for an orphan receptor is later discovered, the receptor is referred to as an "adopted orphan".

Adopted Orphan

↓

Identifying the endogenous ligand for the orphan receptor.

- Classification of known and orphan (non-olfactory) GPCRs -
39% Class I, knowns (Rhodopsin class)
6% Class II
37% unknowns.

- Identification of oGPCRs - The identification of active ligands at orphan GPCRs has been achieved through the development of a no. of experimental approaches such as -

↓ Screening of putative small molecule and peptide ligands

↓ Reverse pharmacology

↓ Use of bioinformatics to predict candidate ligands.
(1) Structure based Virtual Screen
(2) Ligand based Virtual Screen.

* Classical pharmacology / forward pharmacology -
(Ntk)

- Using the performance of medicinal chemistry, the potency, selectivity and other possessions of these screening hits are optimized to create candidate drugs.
- Compounds are screened in cellular or animal replica of infection to categorize compounds that cause a desirable transformation in phenotype.

Target selection / identification → Library screening → Hits / leads → Leads optimization → HITS

(Ntk)

- ..... disease modifying and screen for compounds that modulate the activity of this purified target.
- Later on compounds are hardened in animals to see if they have the desired effect.
- This approach is known as "reverse pharmacology" or "target based drug discovery" (Ntdd).

PAGE No.
DATE / /

* Reverse Pharmacology :-

(MK)

- The identification of a receptor sequence is followed by the discovery of the corresponding ligand.
- Subsequently the pharmacological and physiological context is investigated for eg, by gene knockout or over-expression of the receptor and its ligand.
- This method became popular after the sequencing of the human genome which allowed rapid cloning and synthesis of large quantities of purified proteins.

* Scope :-

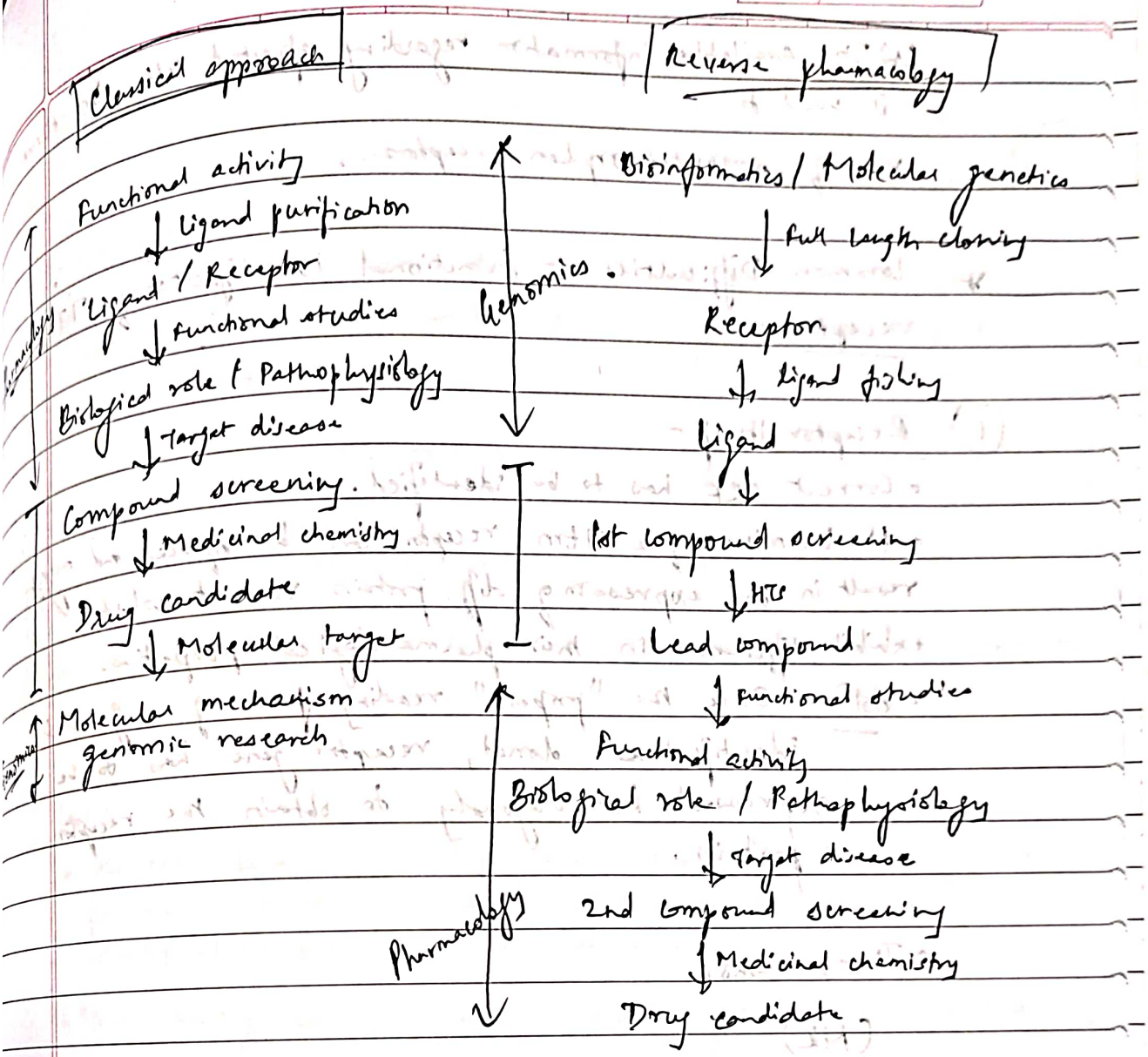
- Understand the MOA at multiple levels of biological organization.
- To optimize safety, efficacy and acceptability of the leads in natural products, based on relevant science.

* Keywords :-

(1) Drug Target – Cellular / molecular structures involved in the pathophysiology of interest where ... (MK)

(2) Hit - potential activity at a chosen target.

(3) Lead - increased activity at a chosen target. (potential)
reduced —" — unrelated target.



☆ Reverse Pharmacology Methods - (★ Diagram for all 3 methods)
(Deorphanization)

① Library based approach -
(NK)

② Tissue extract based -
(a) Isolating active components from crude tissue extract at laboratory scale.
(b) Identifies hits in fractionated tissue samples.

③ Information based -
(a) Identifies prospective ligands by database screening and testing them on a small no. of selected orphan receptors.

(b) The available information regarding affected cells or tissues is used to ...

no. of suspected orphan receptors ...

* Common Difficulties — functional identification of 7TM receptor —

(1) Receptor itself -

- Correct ORF has to be identified.
- N-terminus of a 7tm receptor can be spliced and might result in the expression of diff. protein variants which exhibit differences in their pharmacological properties.
- Ofn — Once the "proper" reading frame of a gene is identified and cloned, receptor gene has to be expressed heterologously to obtain the receptor protein.

- Thus, recombi.

(NK)

For eg —

(a) The receptor can be tagged with either a short peptide (on the N-terminus).

(b) However both of the techniques could potentially influence receptor protein targeting, Disturb ligand binding, or alter the signaling behavior of the receptor.

(c) Hence, the cell surface expression of the receptor has to be controlled in functional assays.

(NK)

(c) the over-expressed receptor might also attract G-Proteins from other endogenously expressed receptors, thus decreasing signals obtained from these receptors.

(d) None of the techniques described is straightforward; all require a considerable amount of work, yet, still they do not guarantee success.

(2) (NK) related to the receptor protein might arise from its possible interaction with ...

(3) coupling of the receptor to effectors inside the cell.

- The difficulty related to receptor signalling & assay systems are shared by all 3 approaches.

(NK)

- However, because of the high demands placed on the assay used for the initial characterization, techniques successfully used for identification purposes are not as numerous.
- These assay techniques have to display high reliability and quality, a property that is defined by a high Z-factor.

Summary

(2) The tissue-extract-based approach -

- Requires an ~~extract~~ extraordinarily sensitive assay procedure ...

Case Study - GPCR

{Reverse Pharmacology of orphan GPCRs}

(NK)

- Later, when high-throughput screening technology was applied to reverse pharmacology, dozens of orphan GPCRs become deorphanized.
- Furthermore, novel neuropeptides were discovered.

PAGE No.
DATE / /

- Artemisinin -

(NK)

- Artimisinin and its derivatives offer promise as a new class of antimalarial
- Best case for reverse pharmacology approach.
- Anti-cancer, asthma and anti-oxidant properties.

(NK)

- Although molecular biological and bioinformatics techniques made the identification of orphan GPCR's responsive, the search for their endogenous ligands has been a challenge.

- This search has given birth to the reverse pharmacology approach, which uses orphan GPCRs as targets to identify endogenous ligands.
- This approach was very successful and has led to over 2 decades to the deorphanization of about 300 GPCRs.

- Searching for Novel Transmitters -

(NK)

- These transmitters are mostly small molecules although few of them are larger polypeptides and they include, neuropeptides, lipid mediators, nucleotides, amino acid and derivatives etc.
- So how could one find which one of these neurotransmitters may bind an orphan GPCRs ?

(NK)

- (NK)

- Because its phylogenetic analysis classified ~~opioid~~ opioid receptor like 1 as a peptidergic (chemical which functions to directly modulate the peptides of the body or brain). GPCR and because the opioid receptor-like is expressed in CNS, peptidergic brain tissue extracts were prepared, purified and fractionated.

(NK)
- Its structural similarities to the opioid peptides made it an immediate star, yet it has been proven not to bind the opioid receptors.
- Further studies need to be carried out.

- Hurdles of Reverse Pharmacology –
- (NK)
- Majority of the drug discoveries would not have been developed or their development would have been delayed significantly in the absence of the scientific or technical contributions from the pharmaceutical companies.
- (NK)
- Best of public and private sector partners comprising academia and industry should come together to reap significant benefits from these seemingly low fashionable but highly gifted explorations based on traditional knowledge.
- (NK)
- Reverse Pharmacology approaches need to be developed further and optimized as novel means for fast track drug discovery and development of newer, safer and effective drugs.

Introduction to Proteomics

PAGE No.
DATE / /

* Protein Identification with Ab –

(NK)

- Each tip → of the "Y" of an Ab contains a paratope (analogous to a lock). This is specific for one particular epitope (analogous to a key) on an antigen.
- This Architecture allowing these 2 structures to bind together with precision.
- Using this binding mechanism, an Ab can tag a microbe or an infected cell for attack by other parts of the immune system, or can neutralize it directly. (for eg, by blocking a part of a virus that is essential for its invasion).
- Abs can be used to visualize the location of specific proteins within the cell.

* The benefit of detecting Ab is How –

(NK)

- There method used is –

(1) Immunocytochemistry / Immunofluorescence – refers to the visualization of specific antigens in culture cells.

(2) Immunohistochemistry – refers to their visualization in prepared tissue sections.

- Immunofluorescence and immunohistochemistry – use antibodies for detection and localization of proteins and other antigens within biological ~~systems~~ samples.

(NK)

PAGE No.
DATE / /

Common steps -

(1) fixation of the sample onto slides or plates, blocking in order to avoid unspecific signal, detection by the use of one or two antibodies, and analysis.

(2) (NK).

(3) This is an imp. step because immune cells may be ~~present~~ present within the tissue.

(4) Such approach avoids an unspecific response because primary Abs can bind to antibody receptors present in immune cells membrane. (eg - Mast cell).

* Immunofluorescence (IF) - (1) diagram

(NK)

- 1st - one Ab is used which binds directly to the target Ag, ∴ this technique is called DIF

- 2nd : ~~Abs~~ It is called IIF when 2 Ab are used & because the Ab allowing the detection binds to a 1° Ab that recognizes the ~~2°Ab~~ antigen on the slide.

* Conclusion -

- ∴ Necessary to -

(1) permeabilizing reagents which expose the epitope and highlight the localization of antigen because there are some antibodies which cannot go into the nucleus ; ∴ the epitope will not be achieved and no signal will be produced.

(2) A sample incubated only with the 2° Ab to determine non-specific binding sites.

(3) Control slide containing cells that either do not express and/or ...

* Immunohistochemistry —

(MK)

- ... precipitate at the location of the protein or fluorescent detection, in which a fluorophore is ~~conjugated~~ conjugated with the Ab and can be visualized using fluorescence microscopy.
- Can be performed using samples with different previous treatments such as cryo preserved slides or paraffin-embedded tissue slides.

(MK)

(4) Negative controls could be slides previously exposed to Ab-specific specific serum (eg- rabbit's serum if rabbit's Ab is used) or isotype-specific Ig as primary antibody.

(5) Controls are also useful for determination of non-specific binding sites to secondary Ab and optimal primary Ab dilution.

(6) Positive controls should also be included to validate proper activity of the diff. reagents.

(7) IHC can be done also with fluorochromic labeled antibodies.

* Protein Identification —

(1) Edman's Degradation (E.D) —

(MK)

- To solve the problem of damaging the protein by hydrolyzing conditions, Pehr Edman created a new way of labeling and cleaving the peptide. Edman thought of a way of removing only one residue at a time, which did not damage the overall sequencing.

- Procedure -

(1) Add Phenyl Isothiocyanate, which creates a phenylthiocarbamoyl derivative with the N-Terminal.
(2) The N-terminal is then cleaved under less harsh acidic conditions, creating a cyclic compound of phenylthiohydantoin PTH-amino acid.
(3) This does not damage the protein and leaves 2 constituents of the peptide.
(4) This method can be repeated for the rest of the residues, separating one residue at a time.

(CHK) This can be done by denaturing the protein at heating it and adding HCl for a long time.
- This causes the individual amino acids to be separated, and they can be separated by ion exchange chromatography.
- They are then dyed with ninhydrin and the amount of amino acid can be determined by the amount of optical absorbance.
- This way, the composition but not the sequence can be determined.

* Drawbacks -

(CHK)
- The separated peptides can be isolated by chromatography.
- They can be sequenced using the Edman method, because of their smaller size.
- In order to put together all the sequences of the diff. peptides, a method of overlapping peptides is used.
- The strategy of divide and conquer ....

10.30
11-12

$x^2 + y^2 = 0$

PAGE No.
DATE / /

(NK)

- However, this method is limited in analyzing larger sized proteins (more than 100 a.a.) because of secondary hydrogen bond interference.
- Other weak intermolecular bonding such as hydrophobic interaction cannot be properly predicted.
- Only the linear seq. of a protein can be properly predicted assuming the seq. is small enough.

* Shotgun proteomics for Proteome Profile -

(self)

- Assignment - EXPASY

* Protein Microarray -