Luciano Milanesi

ITB-CNR

Istituto di Tecnologie Biomediche

Consiglio Nazionale delle Ricerche

Via Fratelli Cervi, 93 20090

Segrate (Mi) Italy

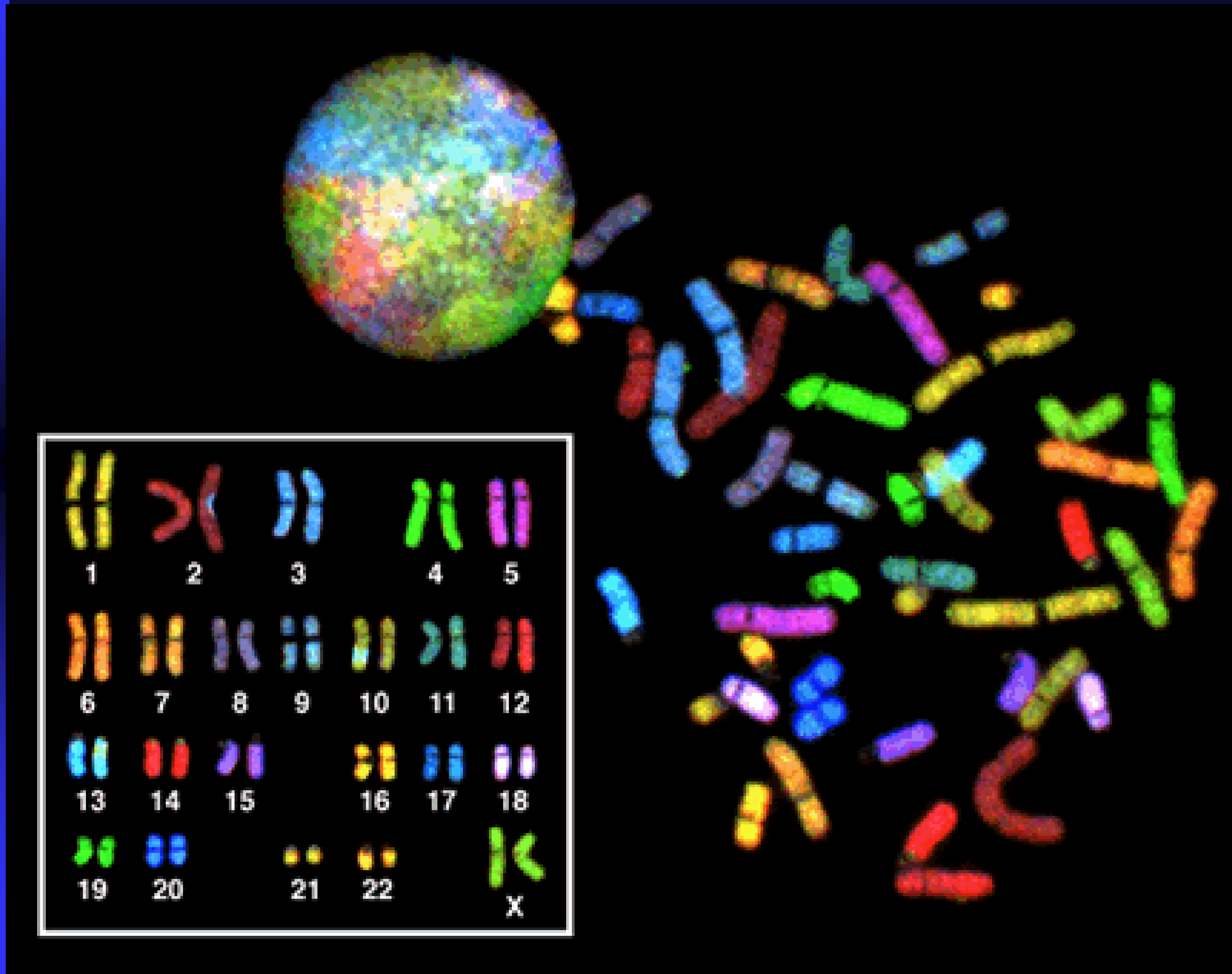
luciano.milanesi@itb.cnr.it

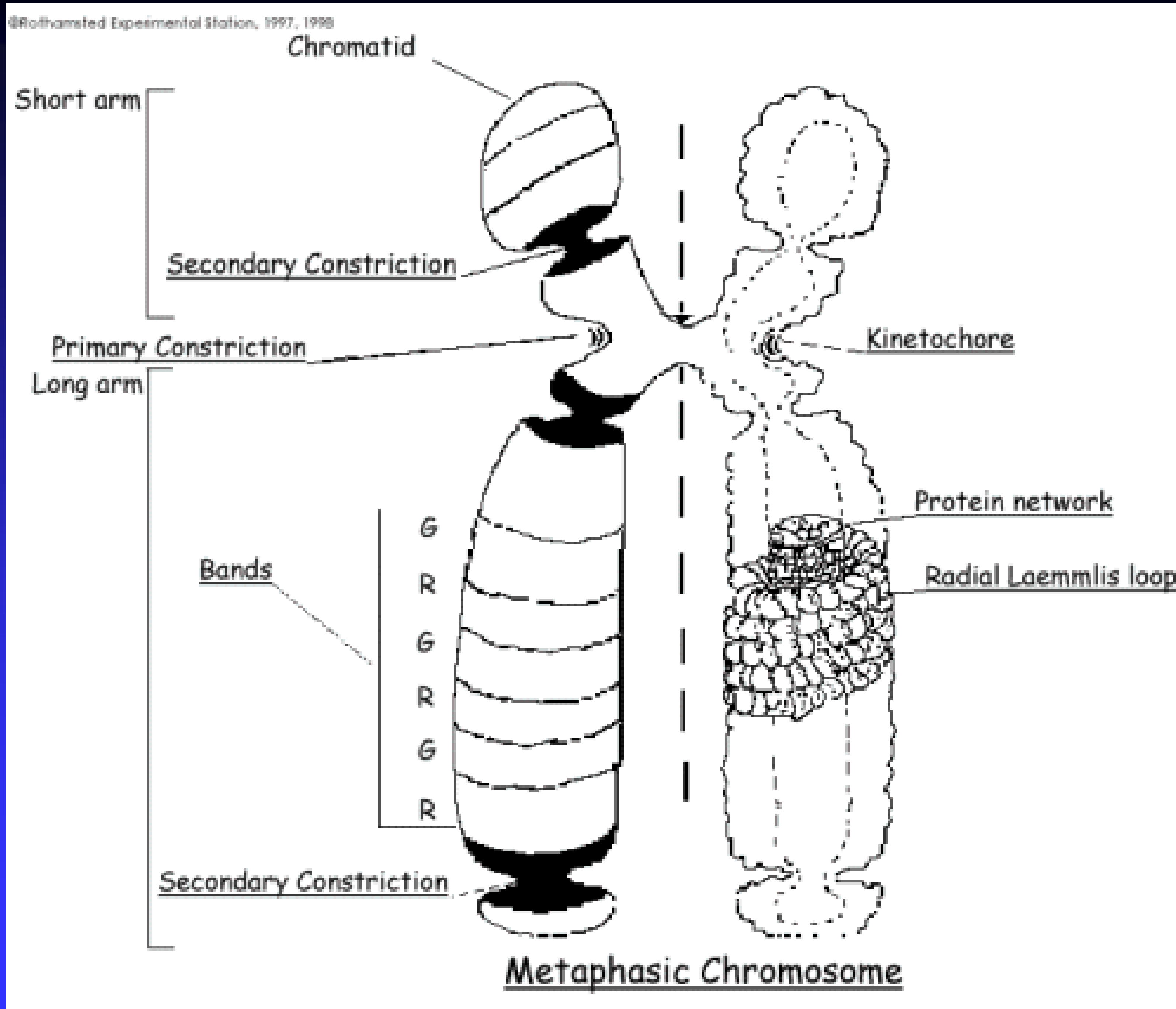
Introduction

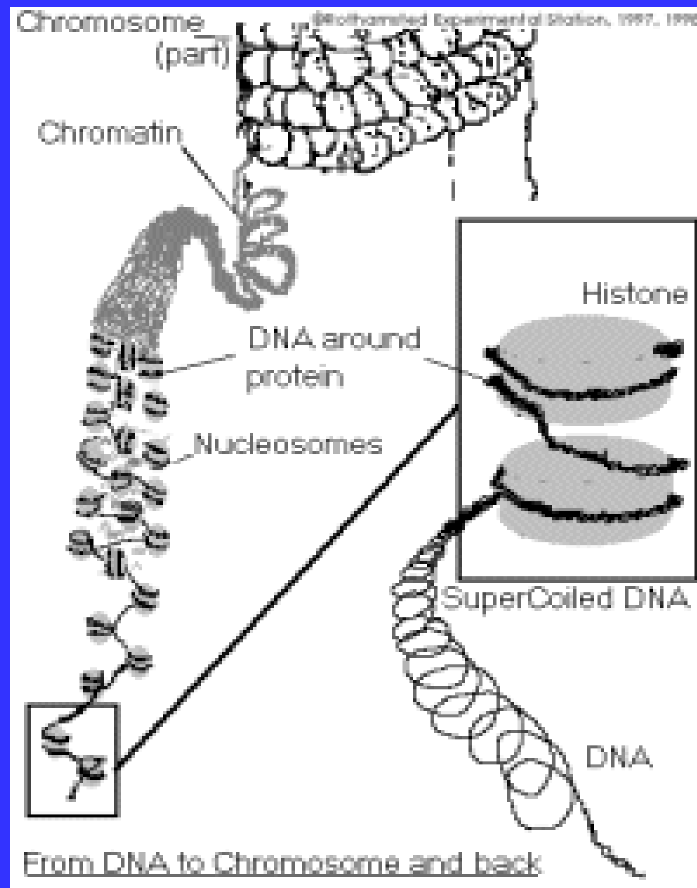
- "Post-genomic" focuses on the new tools and new methodologies emerging from the knowledge of genome sequences.
- Production and use of DNA micro arrays, analysis of transciptome, proteome, metabolome are the different topics developed in this class.

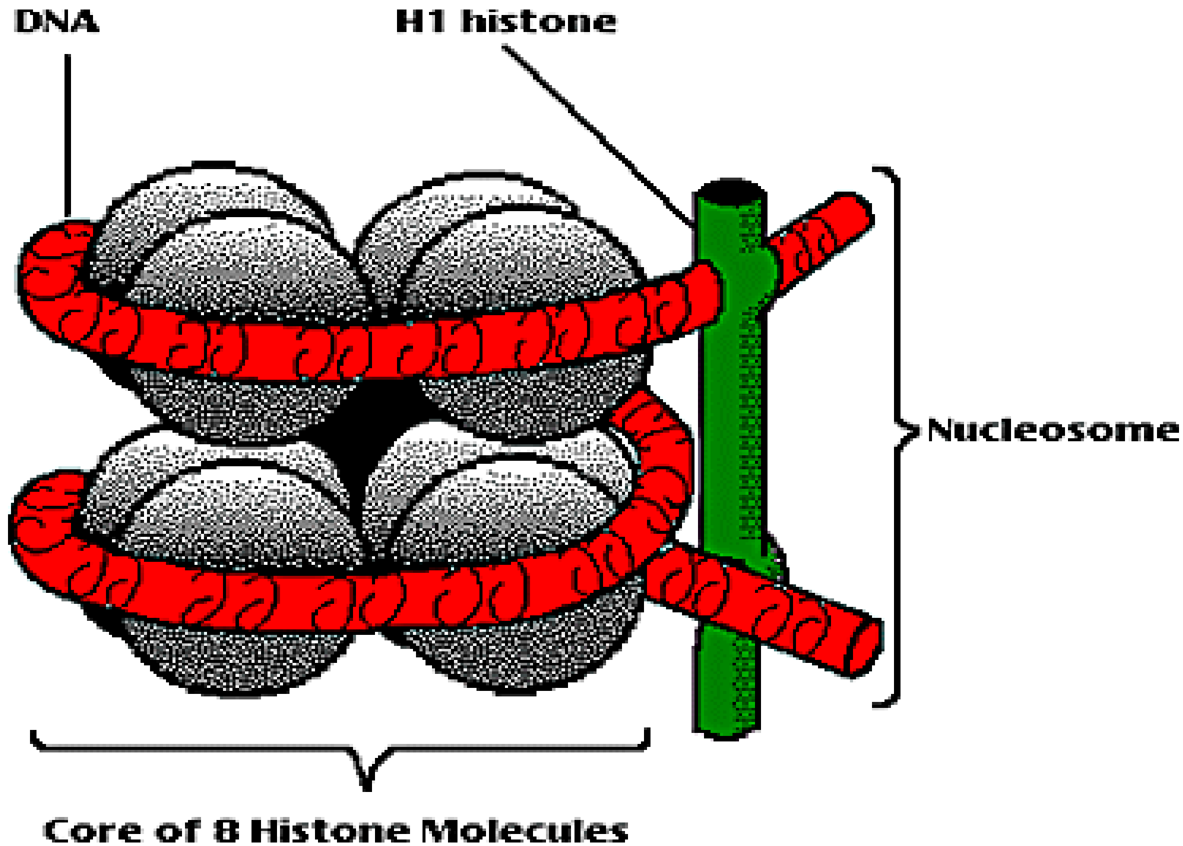
Genome-wide analysis

- Current interest in the genome-wide analysis of cells at the level of transcription ('transcriptome') and translation ('proteome'), the third level of analysis is the 'metabolome'.
- The term 'metabolome' refers to the entire complement of all the small molecular weight metabolites inside a cell suspension of interest.


1 2 3 4 5
6 7 8 9 10 11 12
13 14 15 16 17 18
19 20 21 22 X


©Rothamsted Experimental Station, 1997, 1998
Chromatid
Short arm
Secondary Constriction
Primary Constriction
Kinetochore
Long arm
Protein network
Bands
G
R
G
R
G
R
Radial Laemmlis loop
Secondary Constriction
Metaphasic Chromosome


Chromosome
(part)
Chromatin
Histone
DNA around
protein
Nucleosomes
SuperCoiled DNA
DNA
From DNA to Chromosome and back



Nucleosome

Sequencing Projects Goal

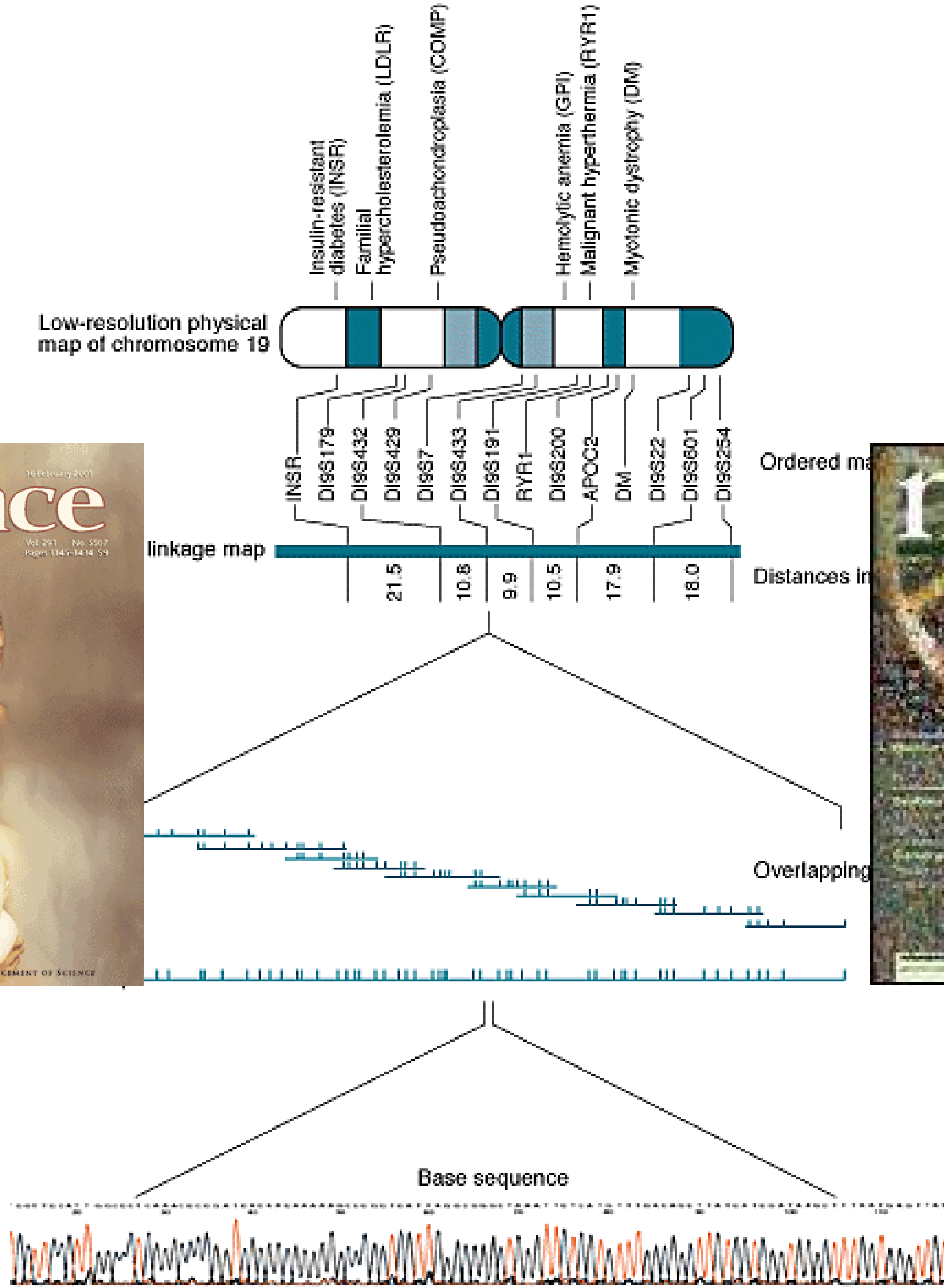
- *identify* all genes in human DNA.
- *determine* the sequences of the 3 billion chemical base pairs that make up human DNA.
- *store* this information in databases.
- *develop* tools for data analysis.
- *address* the ethical, legal, and social issues that may arise from the project.

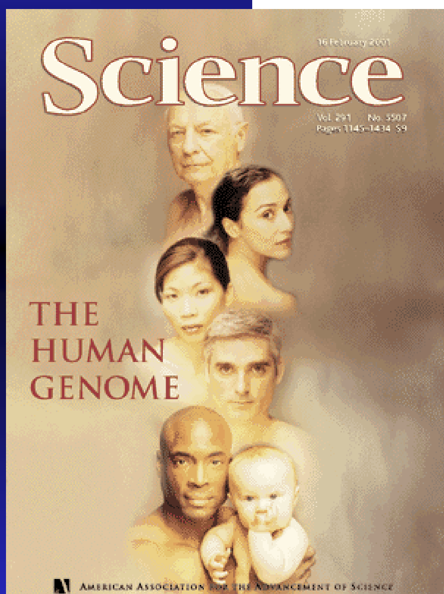
Sequencing Projects Goal

- The ultimate goal of the Human Genome Project and other projects is to sequence the chromosomal DNA of humans and other species and discover the genetic information contained therein.
- The intermediate goal is to construct the physical maps of 23 pairs of chromosome.
- The maps is used for locate feature of interest such as a gene.

Bioinformatics & HGP

- What is the role of Bioinformatics concerning large genome projects like HGP?
- If the full genomic sequence is known, do we know about the functions of each and every gene?
- What is the role of the Internet in Biocomputing?
- Is every piece of information on the Internet useful information?


Insulin-resistant diabetes (INSR)
Familial hypercholesterolemia (LDLR)
Pseudoachondroplasia (COMP)
Hemolytic anemia (GPI)
Malignant hyperthermia (RYR1)
Myotonic dystrophy (DM)
Low-resolution physical map of chromosome 19
INSR
D19S179
D19S432
D19S429
D19S7
D19S433
D19S191
RYR1
D19S200
APOC2
DM
D19S22
D19S601
D19S254
Ordered ma
linkage map
Distances in
21.5
10.8
9.9
10.5
17.9
18.0
Overlapping
Base sequence


Science
THE HUMAN GENOME
American Association for the Advancement of Science


nature
the human genome

Genome Structure

- Protein-coding gene consists of a coding sequence usually interrupted by non coding sequences called introns.
- Introns can also interrupt non-coding regions (introns in 5' and 3' untranslated regions of pre-mRNA).
- The term “exon” is normally applied for regions which are not spliced out from a pre-mRNA sequence (5' untranslated region (5' UTR), coding sequences (CDS) and 3’ untranslated region (3' UTR)). But this term is often used also to indicate the protein-coding regions only.

Genome Structure

- The density of protein-coding and RNA-coding sequences in the human genome is very low .
- The non-coding DNA is usually responsible for the complex regulation of the genome and functioning of genes.
- Differences in the damage and repair rates among different segments of chromatin DNA have been observed.
- The rates of synonymous substitutions along chromatin fibers can also vary significantly from about 2 to about 9 substitutions per nucleotide per 10^9 years (Li and Graur, 1991; Boulinkas, 1992).

Genome Structure & CpG

- The genome primary structure is very heterogeneous. The most well known example of such heterogeneity is CpG islands.
- CpG islands constitute a specific fraction of the genome as, unlike bulk DNA, they are non-methylated and they contain CpG at high frequencies.
- The CpG islands have a significantly higher G+C content compared to the rest of the DNA. CpG islands may be used as gene markers.
- There are about 45,000 CpG islands per haploid human genome (Antequera and bird, 1993).

Repeated Regions

- A large part of the human genome consists of repeated DNA. Repeated DNA can be roughly divided into <u>tandemly</u> and <u>dispersed repeated</u> DNA elements.
- The classical human satellites I-IV (comprising 2-5% of the haploid genome) are made up of tandemly repeated DNA (short reiterated sequences of < 100 nucleotides containing divergent repeats of the pentamer GGAAT).
- The other family of tandemly repeated DNA is the alphoid family of satellite DNA (comprising 4-6% of the haploid genome).

Repeated Regions

- Dispersed repeated elements are presented by different retroposons, which can be divided into viral and nonviral superfamilies based on common structural features.
- The length of nonviral retroposons varies significantly (from 30 up to 6000-8000 nucleotides).
- Almost all short interspersed elements (SINEs) are derived from known RNA polymerase III transcripts such as 7SL RNA and tRNA.
- Long interspersed elements (LINE) are transcribed by RNA polymerase II and contain long open reading frames (ORFs).
- (CCG)n and (CAG)n repeats.

Gene Structure

- The analysis of human genes can not merely be considered as a linguistic analysis of the nucleotide string because the gene structure is made up of many other important features.
- These include:
- Higher-order chromatin structure,
- Non-random nucleosome positioning along the DNA.
- The different features of the three-dimensional structure of the DNA (or RNA).
- Torsion strain on the DNA induced by transcription.

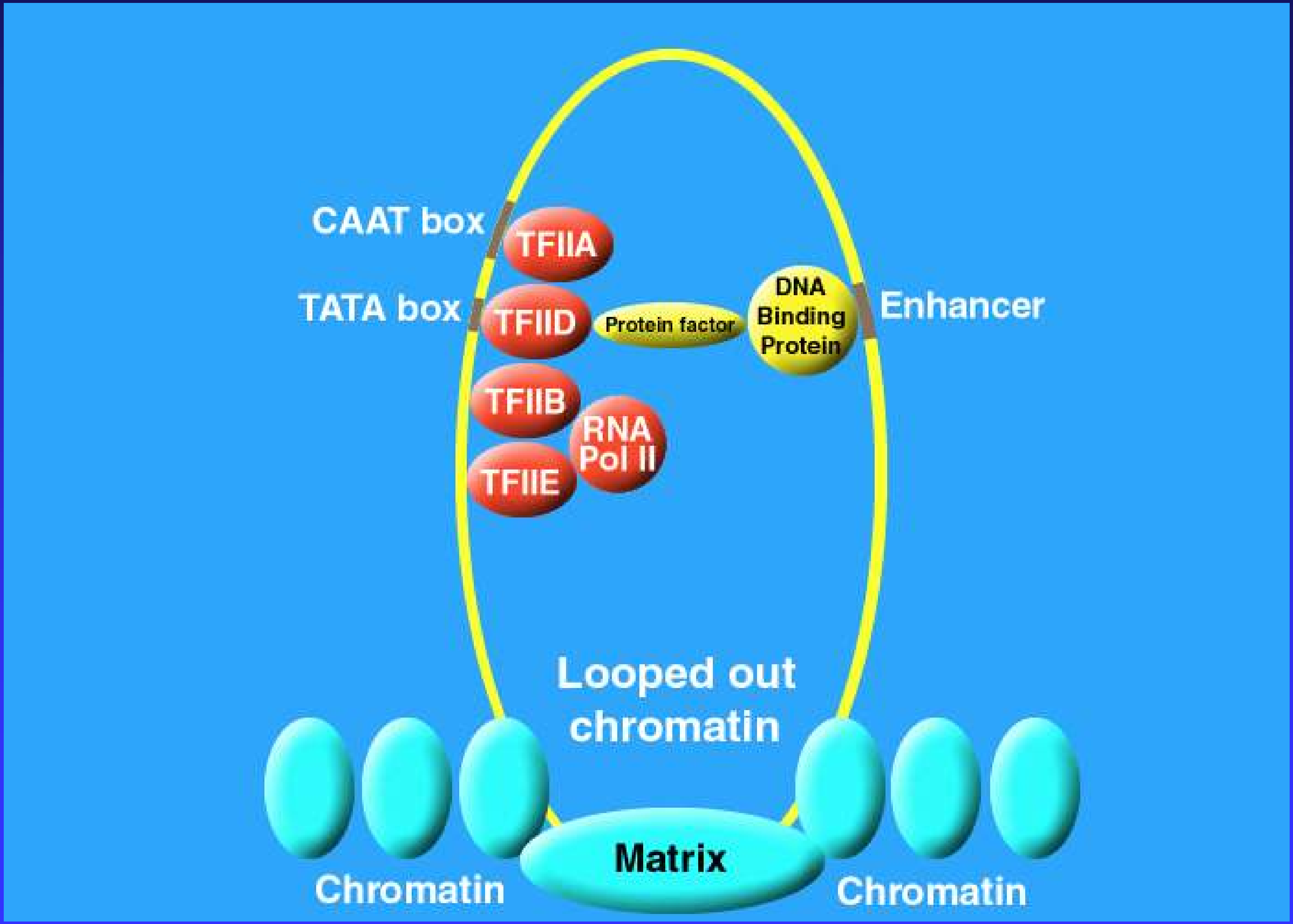
Functional Sites

- The functional unit of a genome is a functional site (also called functional sequence, motif, signal and pattern).
- A combination of functional sites (often binding sites for other molecules) constitutes a functional region.
- The functional unit of a protein-coding sequence is a codon.
- Different functional sites can overlap and/or interact.
- Some other functional sites can be found in protein-coding sequences.
- Other functional units are protein-binding sites.

Functional Sites

- Functional site analysis and recognition is based on the search for general context features common to all sites involved in a specific function.
- Usually, functional sites are recognized by specific factors.
- Invariant structures are very rarely found in sets of functional sites (for example, the AUG initiation site), a large variation is usually typical for functional site structure.

TATA box


CAAT box
TFIIA
TATA box
TFIID
Protein factor
DNA
Binding
Protein
Enhancer
TFIIB
RNA
Pol II
TFIIE
Looped out
chromatin
Chromatin
Matrix
Chromatin

Binding Sites of Transcription Factors

- Protein-coding genes are transcribed by RNA polymerase II. Transcription is initiated in the promoter region by a complex of different factors.
- The elementary sequence signals are typically short (in the range of 5 to 30 base pairs) and highly variable, reflecting the biochemistry of the decoding mechanism.
- Major RNA polymerase II promoter elements are the TATA-box, the cap signal, the **CCAAT-box**, and **GC-box**.
- The **TATA-box** is found in the majority of protein-coding genes. It is usually located about 25 to 30 nucleotides upstream from the transcription initiation site.

Promoter Organization

- Promoters often are responsive to multiple signals.
- Promoters must integrate several signals into one output.
- The arrangement of elements must support 3-D protein complexes.
- Promoter function must be stable against point mutations.
- Promoters must be specifically recognized in the genome.
- All this must be achieved with a limited number of elements.

Promoter Organization

- The arrangement of elements must support 3-D protein complexes.
- The relative order of elements can be crucial.
- The distance of elements is important.

Promoter Organization

- Promoter elements themselves show low sequence conservation.
- All promoter elements can be found almost everywhere in DNA.
- Elements appear not conserved in order and spacing within promoters.
- Sequence similarity between functionally similar promoters can be low.
- Promoters usually are less conserved than the coding parts of genes.

Promoter analysis programs

- All programs concentrate on detection of polymerase II promoters,
- therefore, programs usually work best with short sequences.
- There are two categories of promoter finding programs:
- Programs that find solely core promoters; Specificity is usually very low;
- Search programs that include the proximal promoter;
- Promoter analysis programs cannot find promoters;

Detection of Tata-box in combination with an initiator

- Neuronal network approach (**NNPP**);
- Discriminator analysis (**Zhang's core promoter**);
- Oligonucleotide frequencies (**PromFind, TSSW, TSSG**);
- Correlated word analysis (**CoreInspector**);

Promoter Region

- TF-binding site frequency profiles (**promoter Scan, PromFD**).
- Region-specific frequency profiles (**FunSiteP**).
- Oligonucleotide frequencies (**PromFind, TSSW, TSSG**).
- Word frequency analysis (**XLandscape**).
- Hidden Markow Models (Audic & Claverie).

Specific modeling And/or proximal Promoter region

- TF-binding site & annotation (**TargetFinder**).
- TF-binding site organization algorithm (**ModelInspector**).
- TF-binding site organization / knowledge (**FastM**).
- TF-binding site correlation (**GenomeInspector**).
- TF-binding site organization (**ModelGenerator**).
- Word frequencies / occurences (**XLandscape**).

Promoter analysis Strategy

- **Step 1**: Try to locate a similar sequence in the database by **BLAST** or **FASTA.**

- **Step 2**: Run available promoter finding programs on the sequence:
- **NNPP** (pol II core promoter prediction).
- **PromoterScan II** (full pol II promoter prediction).
- **FunSiteP** (full pol II promoter prediction).
- **TSSG** (full pol II promoter prediction).
- **TSSW** (full pol II promoter prediction).

Promoter analysis Strategy

- **Step 3:** locate potential transcription factor binding sites in promoter, the following programs can be used:
- **Signal Scan**
- **Matrix search**
- **MatInspector**
- **TSSW**
- **TSSG**

Example of ABF1 binding sites

- GTCGTCTCACACG
- ATCTTTGTTAACG
- ATCGTTAATGACG
- GTCACTGTACACG
- GTCACGATATACG
- ATCCCCATTAACG
- ATCTCTCGCAACG
- ATCATTATGCACG
- ATCATTGAAAACG
- GTCGTCTCACACG

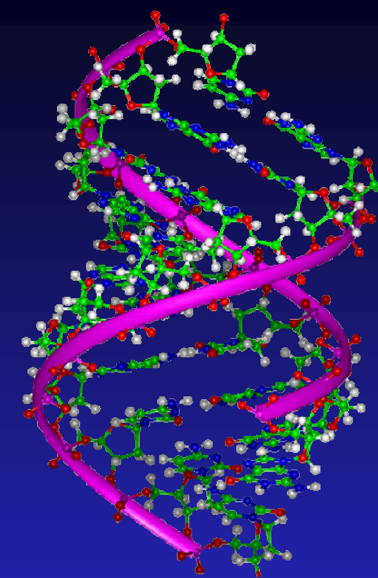
Example of TATA-box weight matrice

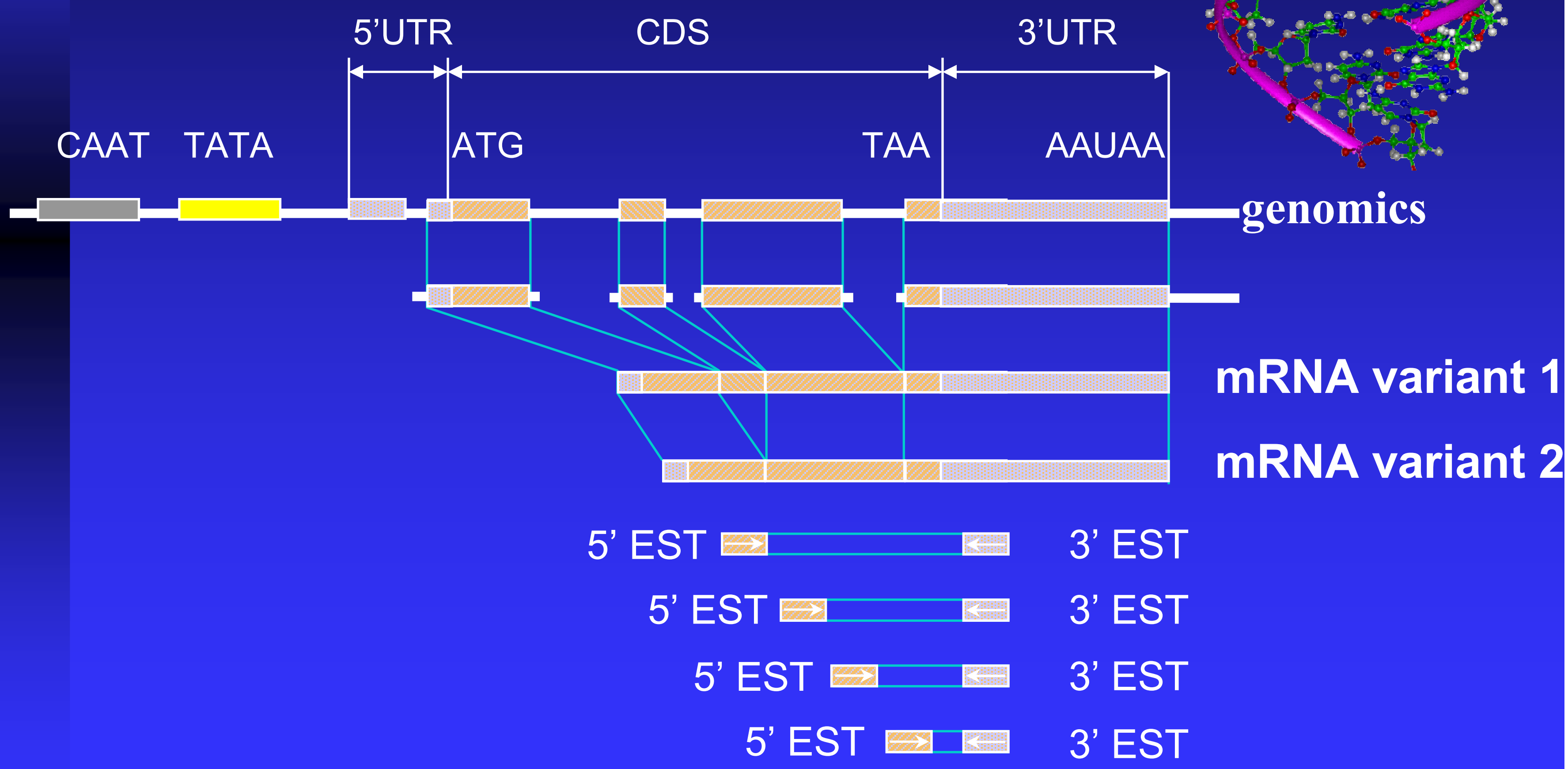
Posi-tions	-3	-2	-1	0	+1	+2	+3	+4	+5	+6	+7	+8	+9	+10	+11
A	16	4	90	1	91	69	92	57	40	14	21	21	21	17	20
C	37	12	0	2	0	0	1	1	11	35	38	33	30	28	26
G	39	5	1	1	1	0	5	11	40	39	33	33	33	36	36
T	8	79	9	96	8	31	2	31	9	12	8	13	16	19	18
	G	T	A	T	A	A	A	A	G	G	C	G	G	G	G
	C		T		T	T		T	A	C	G	C	C	C	C

Gene Prediction

- Prediction of protein-coding genes in newly sequenced DNA becomes very important in large-scale genome sequencing projects.
- This problem is complicated due to the *exon-intron* structure of eukaryotic genes.
- *Introns* are non-coding regions which are spliced out at *acceptor* and *donor* splice sites (Maniatis and reed, 1987).

Gene Structure





Expressed Sequence Tags

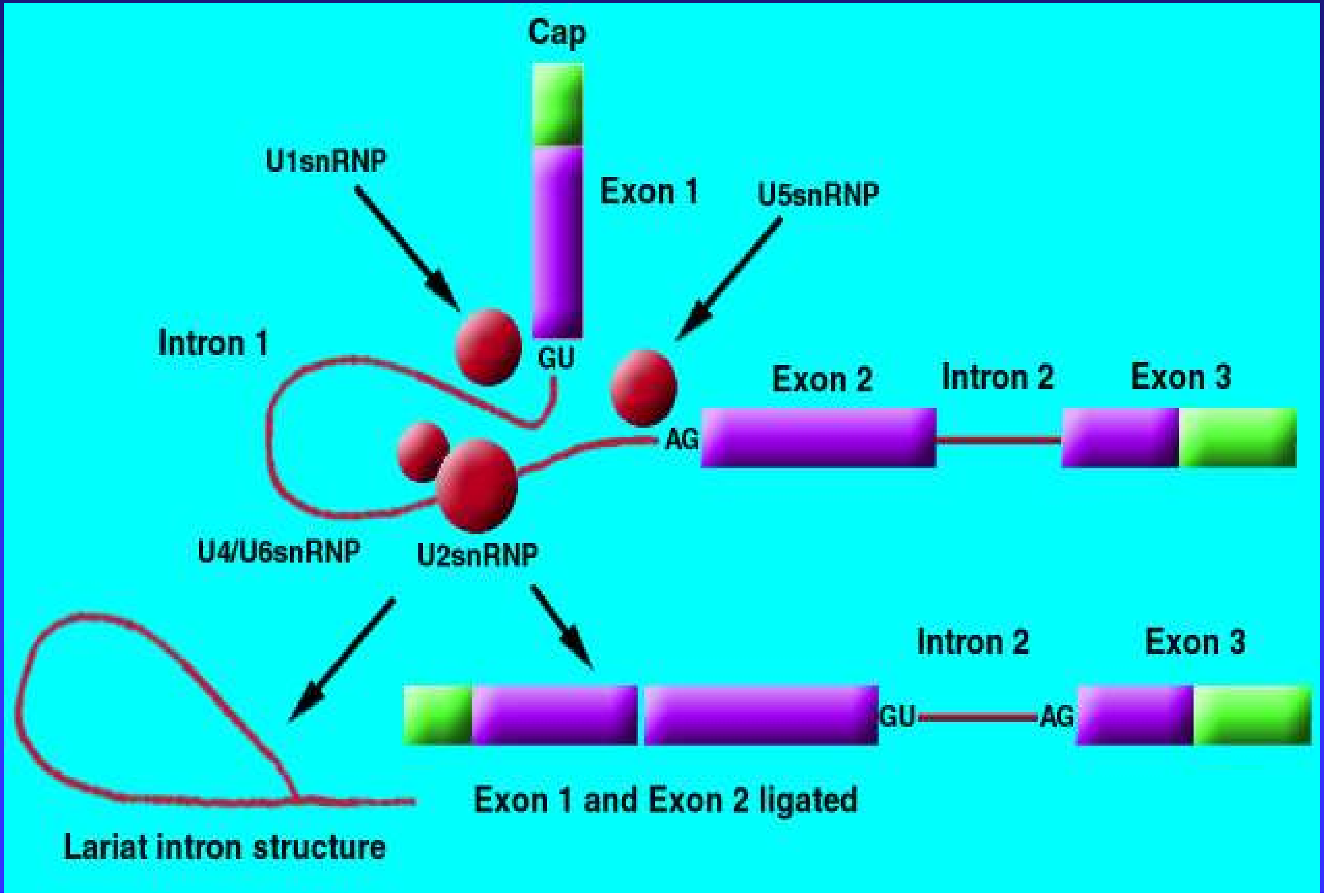
Example of the Initiation Codon ATG

	5′ UTR						CDS		
Posi-tions	-6	-5	-4	-3	-2	-1	+1	+2	+3
	G	C	C	A G	C	C	A	T	G
A	18	19	24	68	23	15	100	0	0
C	21	40	58	2	55	53	0	0	0
G	47	23	12	30	16	23	0	0	100
T	13	18	6	0	7	9	0	100	0

Splice Sites

- Human genes coding for proteins usually include introns and exons.
- Introns are non-coding regions spliced out at <u>acceptor</u> and <u>donor</u> splice sites.
- This process is supported by a complex assembly of snRNP (small nuclear ribonucleoprotein particles) and hnRNPs (heterogeneous nuclear ribonucleoprotein particles), called a <u>spliceosoma</u> (Maniatis and reed, 1987).
- U1 snRNP recognises the donor splice site and U5 snRNP recognises the acceptor one.

Spliceosoma



Polyadenylation Signal

- The most well-known signal involved in this process is AATAAA, located 15-20 nucleotides upstream from the poly-(a) site (site of cleavage) (Proudfoot, 1991).
- About 90% of mRNA have a perfect copy of this sequence. The most frequent natural variant, ATTAAA.
- Analysis of neighboring bases showed that some other bases can be important for AATAAA recognition (milanesi et al., 1996).
- An additional signal with consensus YGTGTTYY (diffusive GT-rich sequence) was revealed in region from 20 to 30 nucleotides downstream of poly-(a) site (site of cleavage).

Gene Prediction Open Problems

- The different methodologies for revealing splice sites and coding regions alone are not able to predict the gene structure.
- The coding region prediction methods:
- Miss most of the short exons.
- Can not reliably define the exon intron boundaries, while the splice site prediction reveals some of real splice sites together with a great number of false sites.
- Can not reliably define the the promoter region.

Gene Structure Prediction

- The main approach for gene structure prediction combines information about ***local functional*** sites (splice sites, initiation codon) together with ***global features*** of coding regions and introns.
- Although there is no direct biological relation between statistical properties of protein-coding regions (at the mRNA translation level) and splicing (at the pre-mRNA processing level), such combinations can strongly improve gene structure prediction.
- Additional information can be obtained from prediction of promoters and the 3' untranslated regions (at the pre-mRNA processing level).

Genomics Sequence Analysis

- *Step 1. Revealing of repeated elements*.
- This is a very important step since the presence of repeated elements can create problems in sequence analysis.
- Long repeated elements contain ORFs which can recognized by gene identification tools as potential genes.
- It is very complicated to interpret database search results, when the output is saturated by a number of highly-scored matches with repeated elements in nucleotide sequence databases.

Genomics Sequence Analysis

- *Step 2. Homology searches .*
- Similarity searches in databases is a very important point in functional mapping, since significant homology with some known functional region is the most obvious sequence landmark.
- In many computational tools database searches are integrated as part of the system.

Genomics Sequence Analysis

- By using the similarity search in EST sequence databases, it is possible to reconstruct the gene structure with high accuracy, although alternative splicing and the presence of other genes in the analyzed sequence should be taken into account.
- Potential errors of sequencing can also be revealed with high accuracy at this step.

Genomics Sequence Analysis

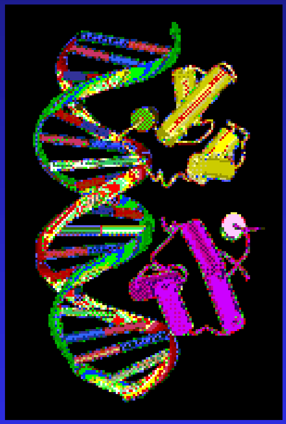
- *Step 4. Gene structure prediction.*
- If no significant homologous protein was revealed (steps 2,3) then the gene structure can be predicted by using coding statistics and potential functional motifs (splicing signals, initiating codons).
- Defining the location of potential splice sites.
- Calculate the potential coding fragments (PCF).
- Selection of a set of "best" PCF.
- Reconstruction of potential gene models.
- Determination of the best gene structure.

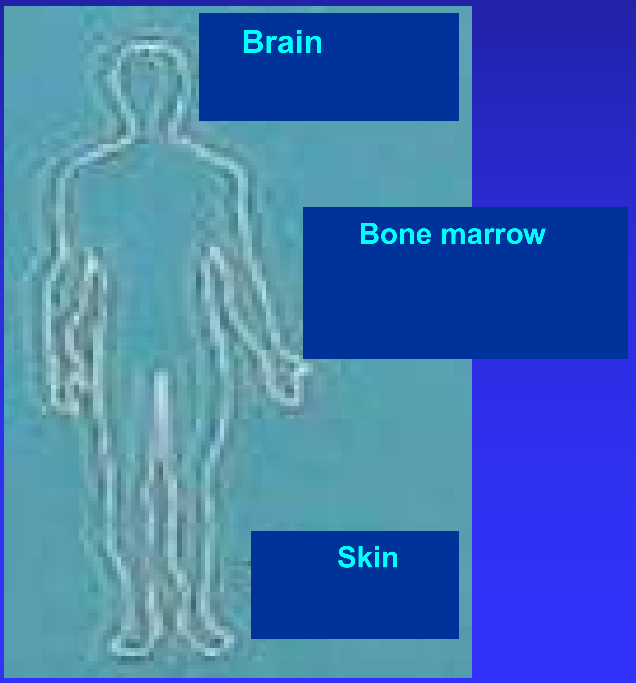
Genomics Sequence Analysis

- *Step 5. Other features*
- The prediction of potential transcription binding sites, promoters and poly(a) signals can help in understanding of the functional meaning of the analyzed sequence.
- Analysis of CpG islands can be very important for gene regions recognition.

Genomics Sequence Analysis

- *Step 6. The analysis of similarity between potential peptides (translated CDS) and the protein databases*
- This is very important for cases of weak, but significant similarities. Such similarities can be lost during a BLASTX search, since all possible translations are used by this program (as a result, sensitivity will be lower).
- Further analysis of protein secondary structure and functional motifs can confirm the full structure of the revealed gene.
- Steps 3-6 can be repeated several times, when more than one gene are present in a sequence.





Brain
Bone marrow
Skin

Functional Genomics

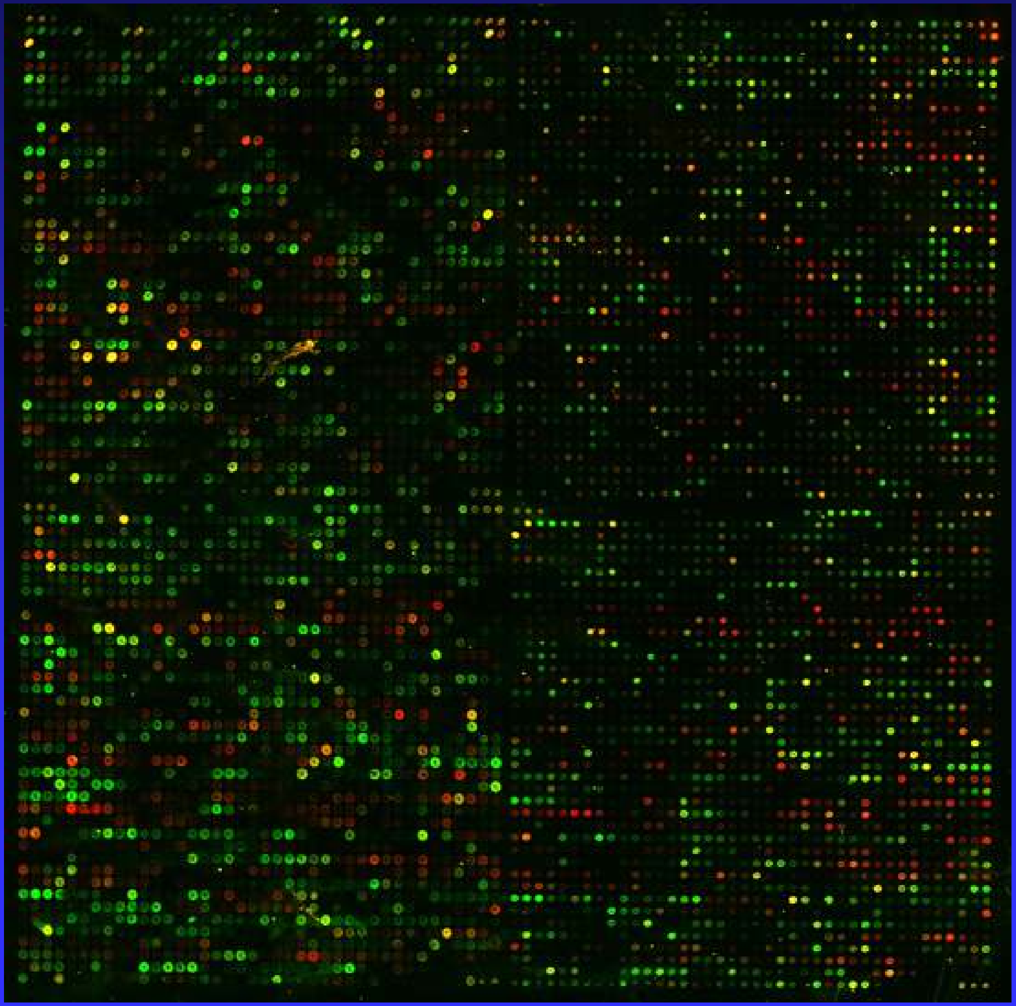
- A new level of systematics experiments is required to obtain an overall picture of When, Where, and How gene are expressed. The study of *functional genomics* includes:
- The analysis of gene expression profiles at the mRNA and protein levels (*Proteome*) and
- The analysis of polymorphism or mutation patterns in the genome (eg. *DNA chips*).

Example Of Microarray Image

Yeast Genome

(6200 genes)

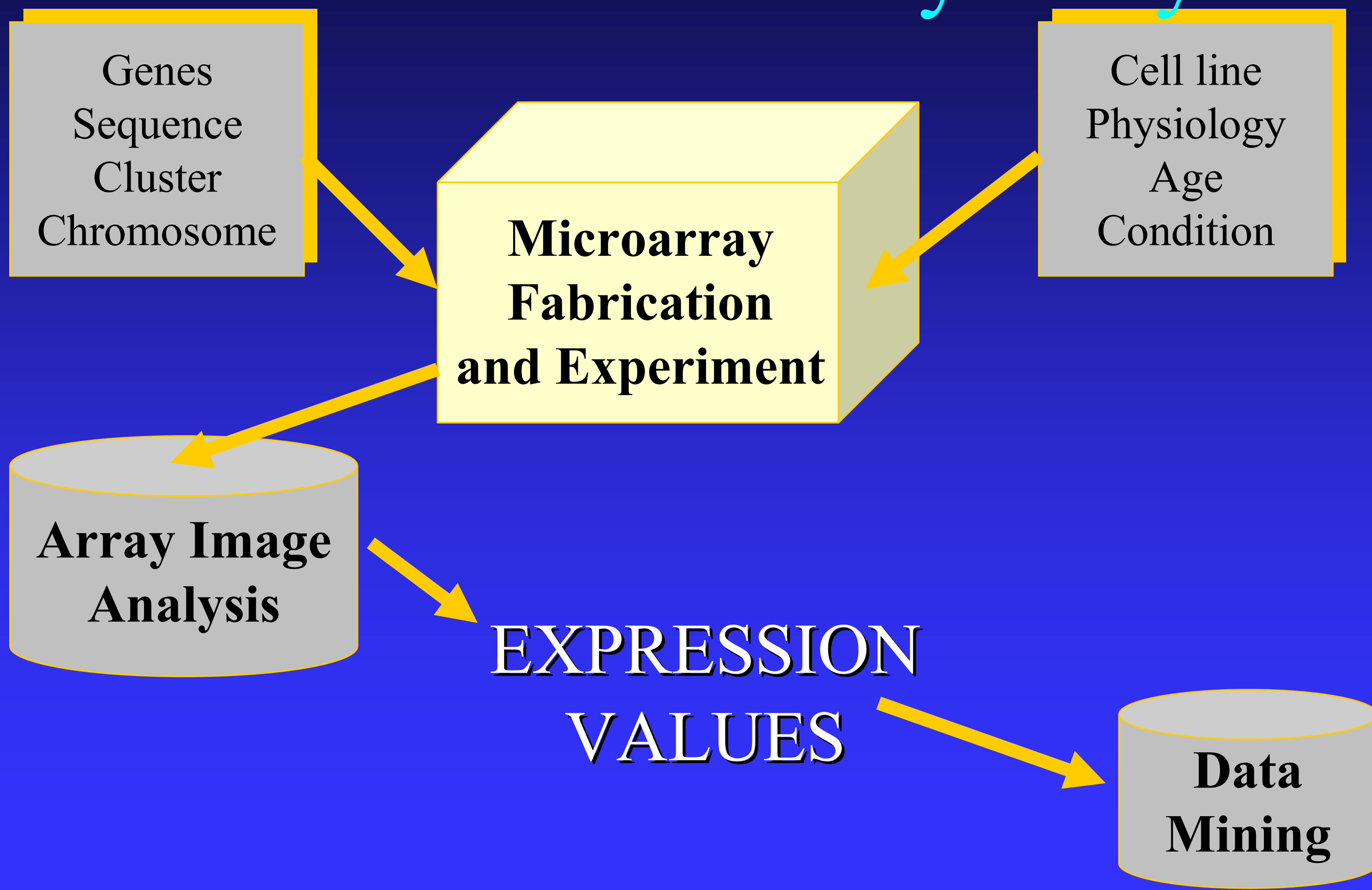
(Brown Lab - Stanford University, Department of Biochemistry, Stanford, CA)



Factors Affecting Expression Profiles

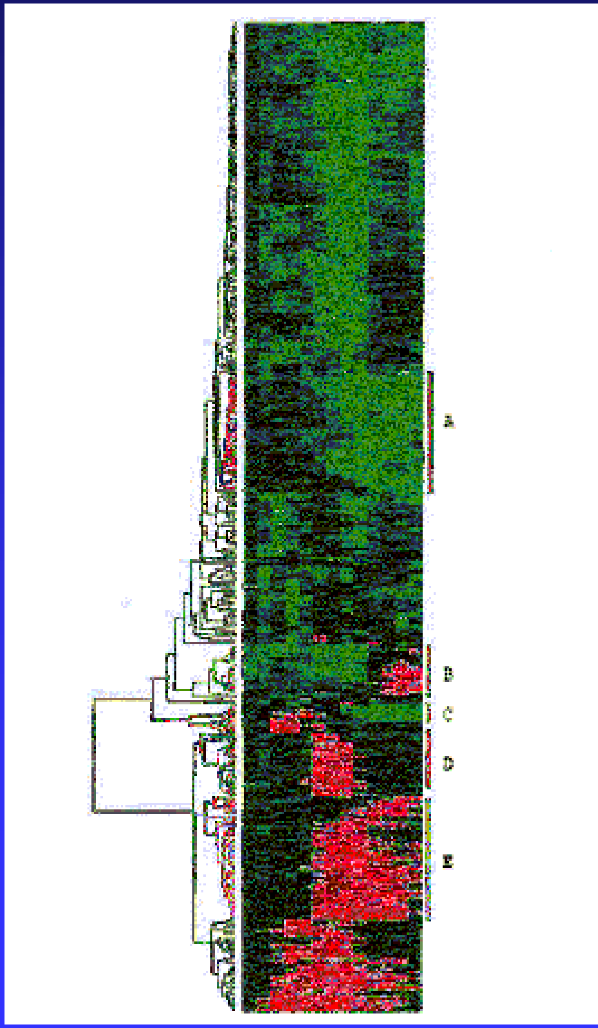
- Every cell owns the entire genetic inheritance of the organism; what makes a cell different from the other ones, is simply its expression profile: which genes are active and how much.
- The expression profile depends on:
- **cell type** (tissue)
- **development stage** (cellular differentiation)
- **health conditions**
- **environment stimuli** (heat, chemical substances, toxic substances, etc.)

Scheme of a microarray assay



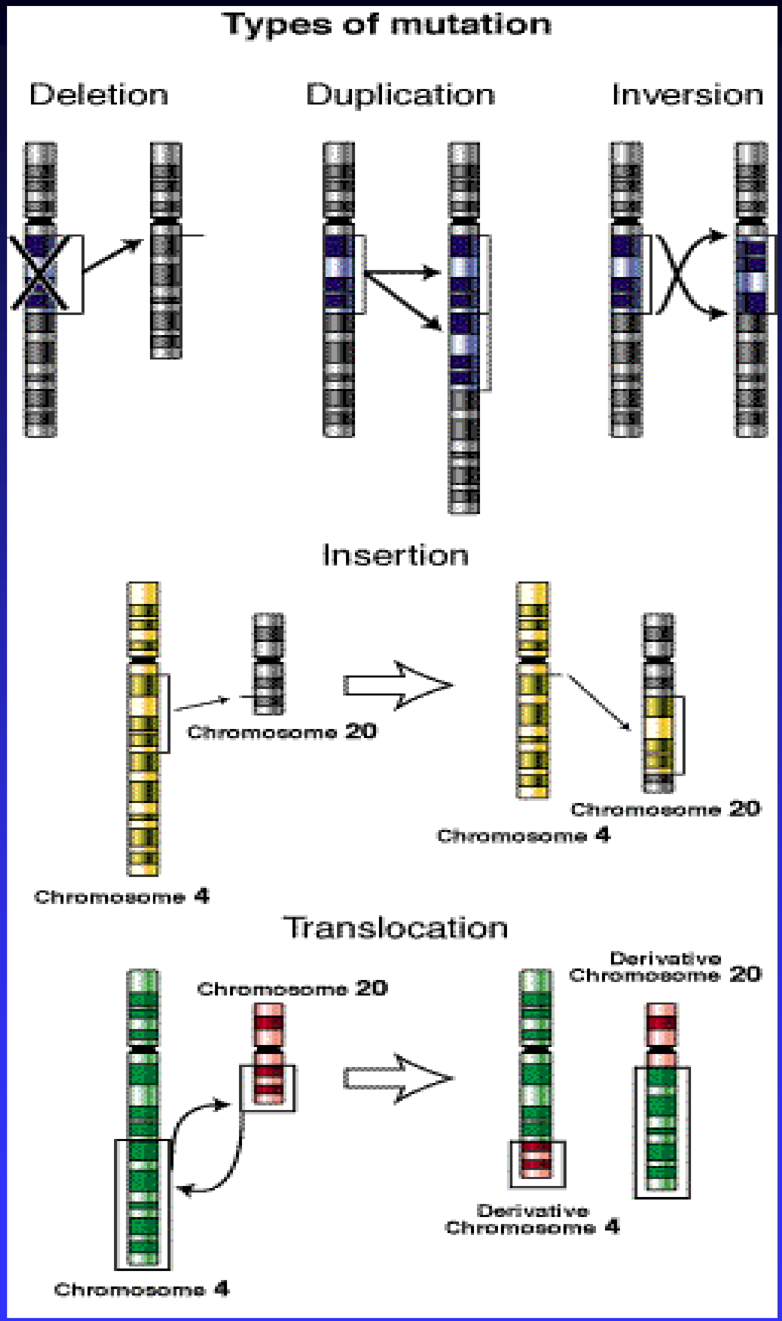
Microarray Data Mining

- Result of a hierarchical clustering. Each row represents a target gene.
- Each expression pattern is visualized by a change in color from green to red.
- Green means repression of the gene with respect to the reference sample, while red induction of the gene with respect to the reference sample.
- Black means no induction and no repression. Color intensities give the magnitude of the phenomenon. Picture from (Eisen et al., 1998).



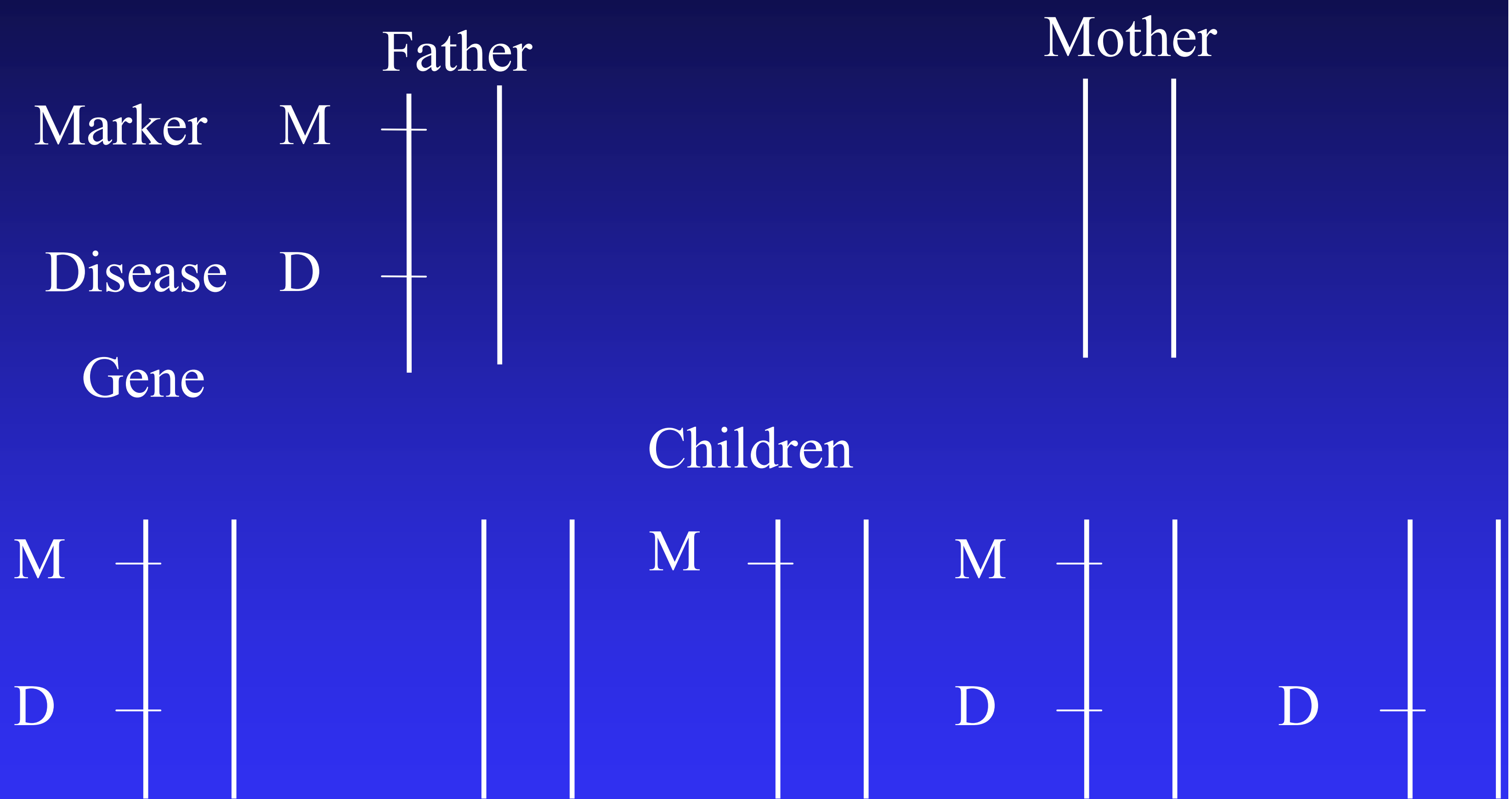
Human Genetic Diversity

- Any two individuals differ in about **3×10^6 bases (0.1%).**
- The population is now about **5×10^9**.
- A catalog of all sequence differences would require **15×10^{15}** entries.
- This catalog may be needed to find the rarest or most complex disease genes.



Types of mutation
Deletion
Duplication
Inversion
Insertion
Chromosome 20
Chromosome 4
Chromosome 20
Chromosome 4
Translocation
Chromosome 20
Derivative Chromosome 20
Derivative Chromosome 4
Chromosome 4

Linkege analysis.



The relative distance between the disease gene (D) and the marker (M) can be estimated from the frequency that both D and M are transmitted together.

Genetic Linkage Data

- Correlating sequence information with genetic linkage data and disease gene research will reveal the molecular basis for human variation.
- If a newly identified gene is found to code for a flawed protein, the altered protein must be compared with the normal version to identify the specific abnormality that causes disease.

Genetic Linkage Data

- Once the error is pinpointed, researchers must try to determine how to correct it in the human body, a task that will require knowledge about how the protein functions and in which cells it is active.
- Correct protein function depends on the three-dimensional (3D), or folded, structure the proteins assume in biological environments; thus, understanding protein structure will be essential in determining gene function

Human Genome and Medicine

- The atlas of the human genome will revolutionize medical practice and biological research into the 21st century and beyond.
- All human genes will eventually be found, and accurate diagnostics will be developed for most inherited diseases.
- In addition, animal models for human disease research will be more easily developed, facilitating the understanding of gene function in health and disease.
- Researchers have already identified single genes associated with a number of diseases, such as:
- **cystic fibrosis, Duchenne muscular dystrophy, myotonic dystrophy, neurofibromatosis, and retinoblastoma.**

Human Genome and Medicine

- As research progresses, investigators will also uncover the mechanisms for diseases caused by **several genes** or by a **gene interacting with environmental factors**.
- **Genetic susceptibilities** have been implicated in many major disabling and fatal diseases including **heart disease, stroke, diabetes, and several kinds of cancer**.
- **The identification of these genes and their proteins will pave the way to more-effective therapies and preventive measures**.
- Investigators determining the underlying biology of genome organization and gene regulation will also begin to understand how humans develop from single cells to adults, why this process sometimes goes awry, and what changes take place as people age.

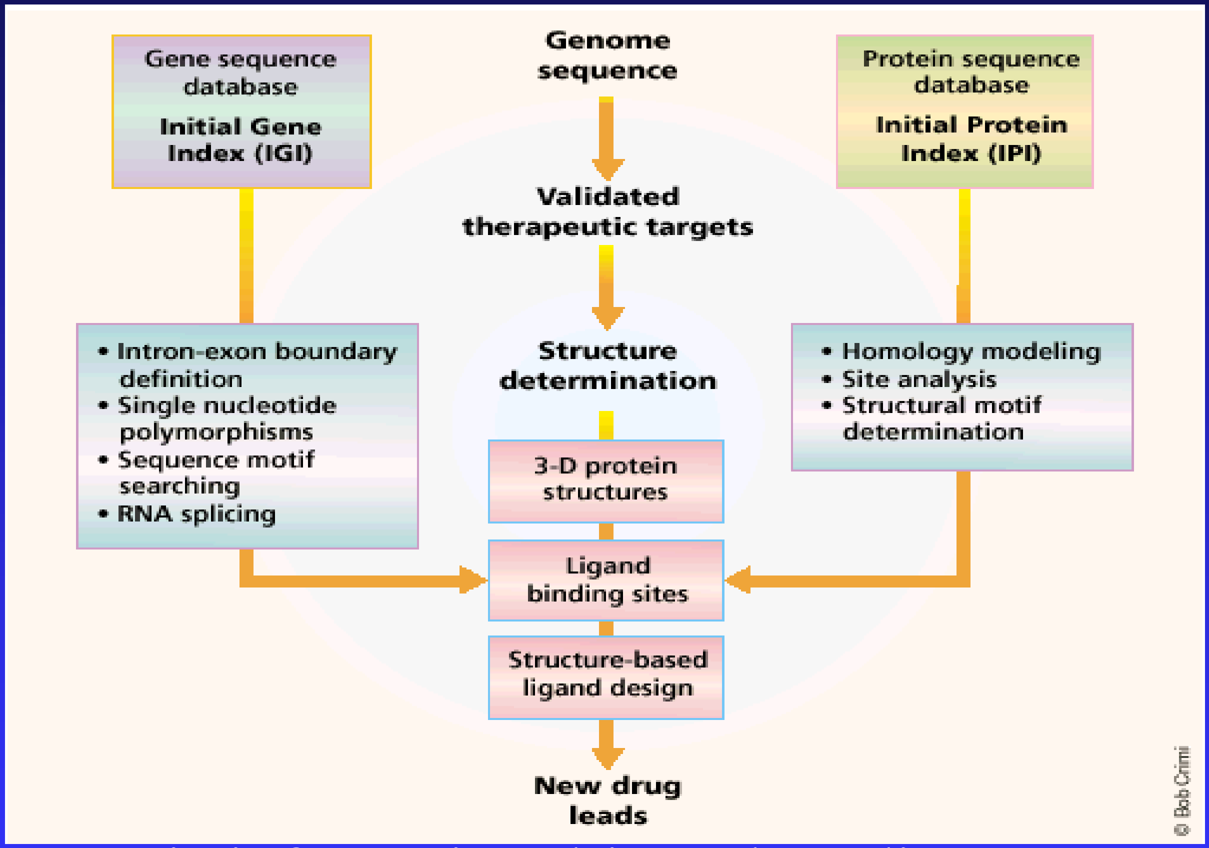
Genomics Medicine

- Less than 5% of the 3.2 gigabase genome encodes genes. The remaining sequence contains a large number of repeats.
- A conservative estimate of the number of genes in the Initial Gene Index gives a value of about 24,500 true genes.
- The structural diversity within the proteins encoded by these genes is considerably greater than this small number of true genes.

Genomic Medicine

- This increased level of protein functionality appears to have been achieved by iterative gene duplication and domain evolution over millions of years, complemented by extensive use of alternative splicing to generate combinatorial diversity at the protein level.
- The structural diversity displayed within the proteome will enable an understanding of the origins and architectures of ligand-binding sites within proteins, a key goal for both structural biologists and drug discovery scientists within the pharmaceutical industry.

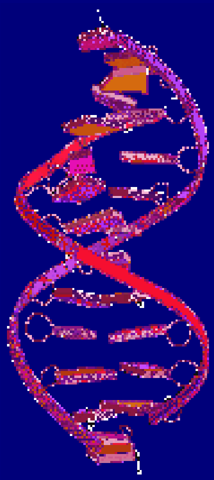
Genomic Medicine



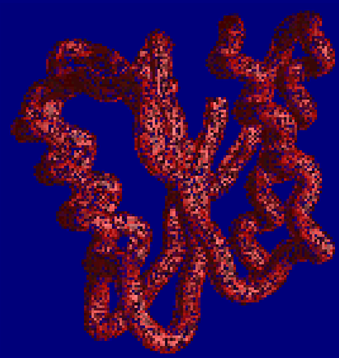
Genomic information-driven drug discovery.
The Initial Gene Index and Initial Protein Index are mined in a variety of ways to identify ligand binding sites.
These in turn can act as templates for structure-based design.

Structural Genomics

Gene



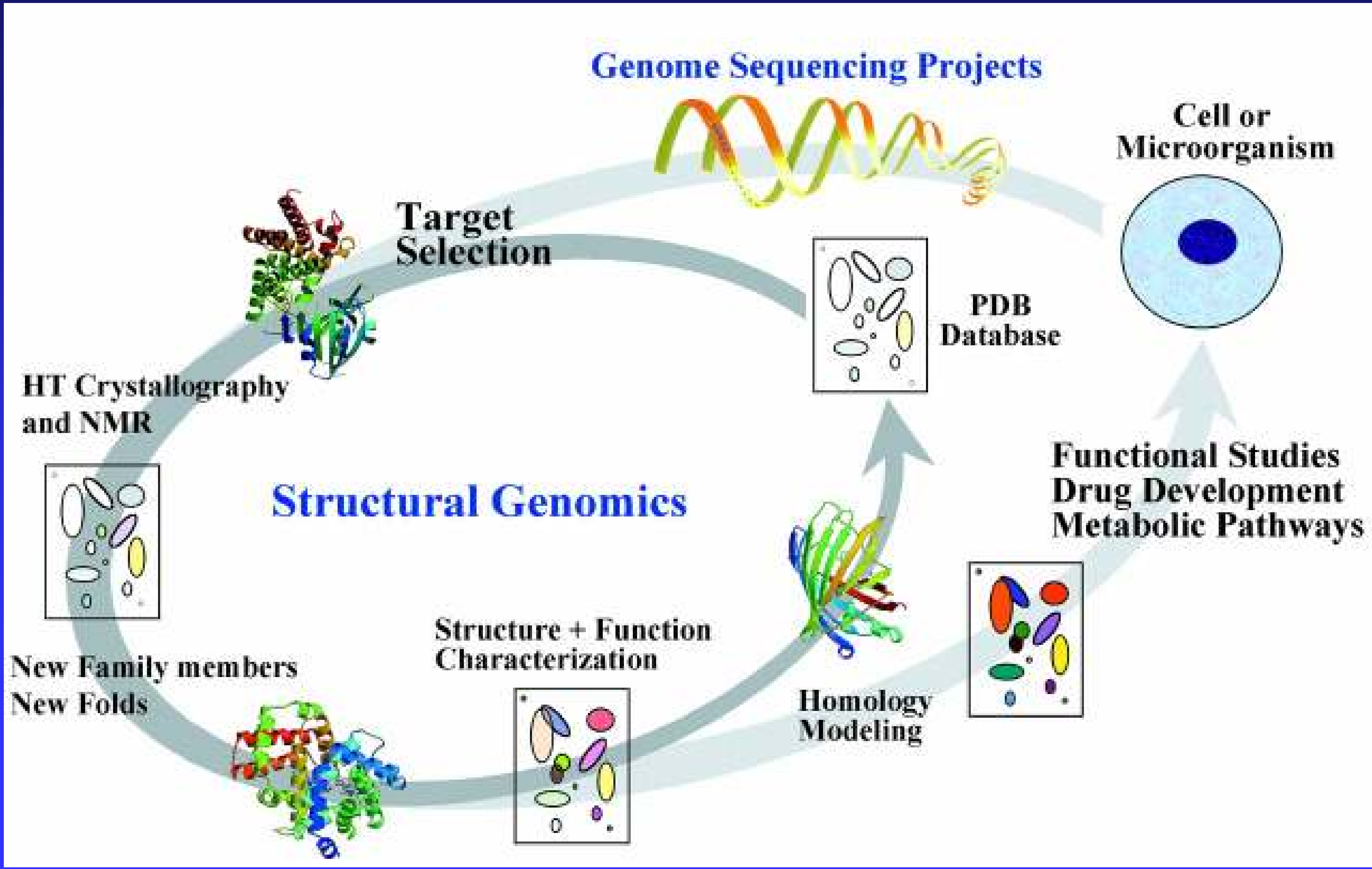
Function



```
> DNA sequence
AATTCATGAAAATCGTATACTGGTCTGGTACCGGCAACAC
TGAGAAAATGGCAGAGCTCATCGCTAAAGGTATCATCGAA
TCTGGTAAAGACGTCAACACCATCAACGTGTCTGACGTTA
ACATCGATGAACTGCTGAACGAAGATATCCTGATCCTGGG
TTGCTCTGCCATGGGCGATGAAGTTCTCGAGGAAAGCGAA
TTTGAACCGTTCATCGAAGAGATCTCTACCAAAATCTCTG
GTAAGAAGGTTGCGCTGTTCGGTTCTTACGGTTGGGGCGA
CGGTAAGTGGATGCGTGACTTCGAAGAACGTATGAACGGC
TACGGTTGCGTTGTTGTTGAGACCCCGCTGATCGTTCAGA
ACGAGCCGGACGAAGCTGAGCAGGACTGCATCGAATTTGG
TAAGAAGATCGCGAACATCTAGTAGA
```

```
> Protein sequence
MKIVYWSGTGNTEKMAELIAKGIIESGKDVNTINVSDVNI
DELLNEDILILGCSAMGDEVLEESEFEPFIEEISTKISGK
KVALFGSYGWGDGKWMRDFEERMNGYGCVVVETPLIVQNE
PDEAEQDCIEFGKKIANI
```

Structural Genomics



Structural Genomics

- Rapidly growing integrated and specialized data bases (e.G. Data bases of mutations, diseases, metabolic pathways, protein structures, etc.) greatly increase the possibility of functional investigation, provided that they are used in a single Information Technology environment.
- In silico comparative analysis of genes can be used in cell cycle for structural motifs determination in the homologous protein families.

Structural Genomics

1. SWISS-PROT and TrEMBL protein sequence databases
2. InterPro - integrated resource of protein families, domains and sites.
3. CluSTr - A database of clusters of SWISS-PROT+TrEMBL proteins.
4. HSSP and PDB for mapping of structure to proteins in each proteome
5. Gene Ontology (GO)

Structural Genomics

- Currently, bioinformatics algorithms, such as PROSITE, can identify around 1,400 distinct sequence patterns, based on their occurrence within linear protein sequences.
- Efforts will continue to focus on developing methods that expand this search to three-dimensional protein architectures, allowing the comprehensive definition of all possible ligand-binding sites.

Structural Genomics

- A key development in the computational world has been the arrival of *de novo* design algorithms that use all available spatial information to be found within the target to design novel drugs.
- Coupling these algorithms to the rapidly growing body of information from structural genomics provides a powerful new route for exploring design to a broad spectrum of genomics targets, including more challenging examples such as protein–protein interactions.

Structural Genomics

- Sequential waves of technology innovation in genomics, proteomics, and drug design will transform the way in which we find new medicines.
- With the genome sequence now in hand, it is clear that such processes as gene duplication and splicing have generated staggering diversity at the protein level.
- This will require a new wave of proteomics technologies.

Structural Genomics

- Three-dimensional information on protein structure emerging from industrial-scale crystallographic approaches will be combined with increasingly powerful *in silico* approaches to allow the design of new drug candidates.

Structural Genomics

Table 1. Three waves of innovation in drug discovery

Innovation	Tools/technologies	Representative companies
First wave		
Genomics (target discovery, antisense therapeutics, and ultimately gene therapy)	Gene / genome sequencing and expressed sequence databases	Incyte (Palo Alto, CA), Celera (Rockville, MD), Genset (Paris) Genome Therapeutics (Waltham, MA)
	Full-length cDNAs/expression	Clontech (Palo Alto, CA), Stratagene (La Jolla, CA), Life Technologies (Rockville, MD)
	Functional genomics	Curagen (New Haven, CT), Millennium (Cambridge, MA), and Pharmagene (Cambridge, UK)
	RNA expression profiling	Affymetrix (Santa Clara, CA), Gene Logic (Gaithersburg, MD), Rosetta Inpharmatics (Kirkland, WA)
	Transgenesis / directed mutagenesis	Lexicon Genetics (Woodlands, TX), Genome Systems (St. Louis, MO)
	Antisense Technology	Isis Pharmaceuticals (Carlsbad, CA), Hybridon (Cambridge, MA)
	SNP databases	Variagenics (Cambridge, MA), Genaissance (New Haven, CT), Celera
	Genetic mapping / disease genes	Oxagen (Boulder, CO), Myriad (Salt Lake City, UT), Decode (Reykjavik, Iceland)
	Genetic diagnostics	Quest (Terterboro, NJ), Roche (Nutley, NJ), Diadexus (Palo Alto, CA), Affymetrix
Second wave		
Proteomics (target validation, drug screening, antibody therapeutics, and protein therapeutics)	Protein databases and protein expression analysis	Oxford Glycosciences (Oxford, UK), Large Scale Biology (Vacaville, CA), Proteome (Cambridge, MA)
	Protein expression technologies and protein therapeutics	Amgen (Thousand Oaks, CA), Genentech (S. San Francisco, CA) Human Genome Sciences (Rockville, MD), Chiron (Emeryville, CA), Genetics Institute (Cambridge, MA), Lonza (Slough, UK)
	Directed evolution	Maxygen (Redwood City, CA), Phylos (Lexington, MA),
	Antibody engineering	Celltech/Medarex (Leatherhead, UK), Abgenix (Fremont, CA) Cambridge Antibody Technology (Cambridge, UK)
	High-throughput screening	Aurora Biosciences (La Jolla, CA) Cambridge Drug Discovery (Cambridge, UK)
	Protein interaction databases	Proteome, MDS-Proteomics (Blainville, QC, Canada), Celera
	Affinity selection	Neogenesis (Cambridge, MA), MDS-Proteomics
	Protein pathways / protein chips	Zyomyx (Hayward, CA), Combimatrix (Seattle, WA), Ciphergen (Palo Alto, CA), Sense Proteomics (Cambridge, UK)
Third wave		
Molecular design (protein therapeutics, antibodies, and small molecules)	Protein structure determination	Structural Genomix (San Diego, CA), Syrrx (San Diego, CA), Astex (Cambridge, UK),
	Protein homology modeling	Structural Bioinformatics (San Diego, CA), Geneformatix (San Diego, CA)
	Protein engineering	Sunesis (Redwood City, CA), Sangamo (Richmond, CA) Cambridge Antibody Technology (Cambridge, UK)
	Structure-based small molecule design	Vertex Pharmaceuticals (Cambridge, MA), De Novo (Cambridge, UK)
	Molecular design tools	Molecular Simulations (San Diego, CA), Tripos (St. Louis, MO)

David Bailey *et al.* Nature Biotechnology 19, 207 - 209 (2001)

Comparative Genomics

- Human genome,
- Mouse genome,
- C.elegans genome,
- Drosophila genome,
- Other genomes

- Can be effectively used to identify families of genes involved in cell cycle.

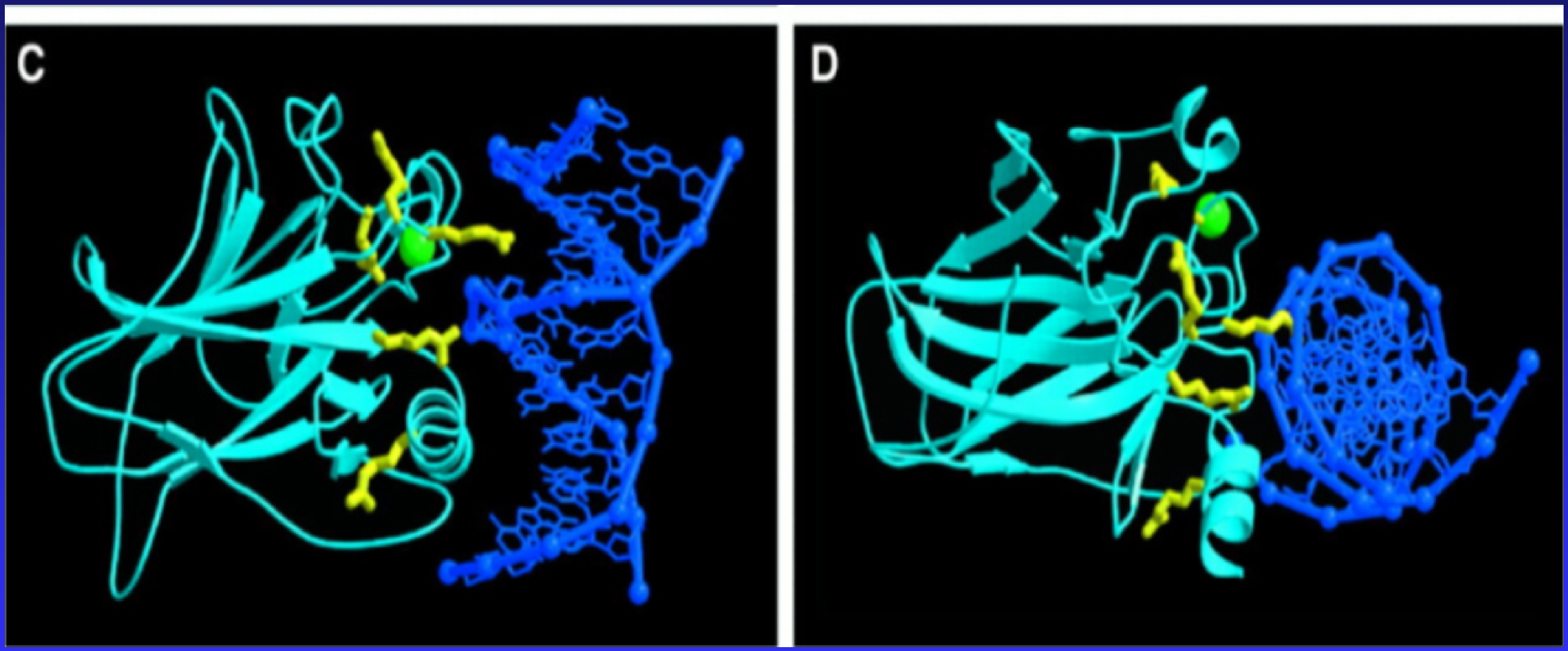
Comparative Genomics

- Homologous gene search in whole genomes. (human, mouse, C.elegans, Drosophila) using. similarity search tools (eg. BLAST, FASTA).
- Gene structure identification : By annotations or prediction (GENSCAN, GeneID, GENEBUILDER, HMMgene ecc.).
- Family grouping of known and predicted genes.

Comparative Genomics

- Sequences of interest can be also annotated in terms of Protein sequence analysis.
- Multiple alignment (eg.CLUSTALW, TCOFFE).
- secondary structure prediction (eg.PredictProtein, PSIpred).
- Pattern detection (eg.PROSITE,BLOCK).
- Subcellular location (eg. Psort).
- 3D homology modelling (eg.Modeller,WHATIF).

Comparison with p53-DNA complex



The site of 53BP2 binding overlaps the site of DNA binding

Gorina S.,Pavletich NP. Science 1996 Nov 8;274 (5289):1001-5

Comparative Genomics

The Bioinformatics analysis will produce

Annotated Sequences from Different genomes

Structural description of interaction interfaces

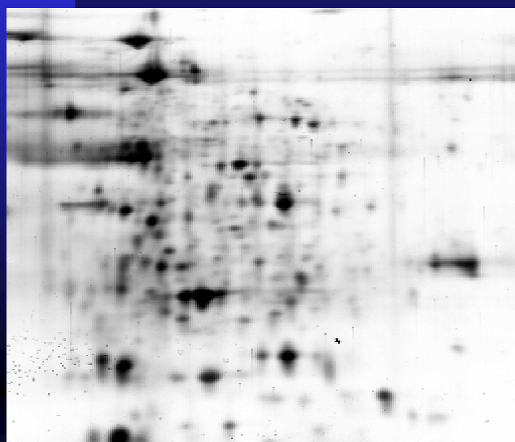
Comparative Genomics

- The informations will be used to obtain: Family assignment of new sequences
- Recognition of structural interaction motifs
- Prediction of interacting proteins
- The identified structural motif can be used to recognize patterns of interaction between a sequence of interest and its interactor
- The predicted complex can be modelled by molecular docking approach (DOCK)

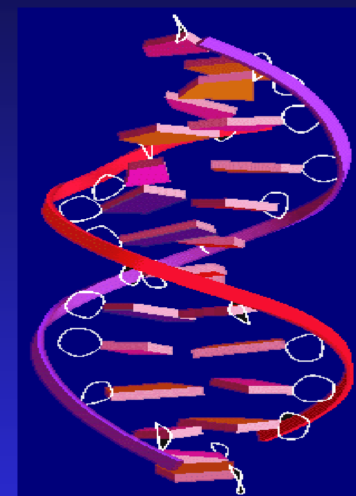
Comparative Genomics

- The informations will be used to obtain:
- Family assignment of new sequences
- Recognition of structural interaction motifs
- Prediction of interacting proteins
- The identified structural motif can be used to recognize patterns of interaction between a sequence of interest and its interactor
- The predicted complex can be modelled by molecular docking approach (DOCK)

Genomics and Proteomics



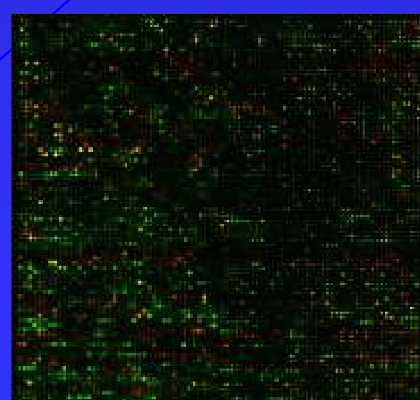
Complex disease mapping



(Proteoma)

Gene & SNPs (Genome)

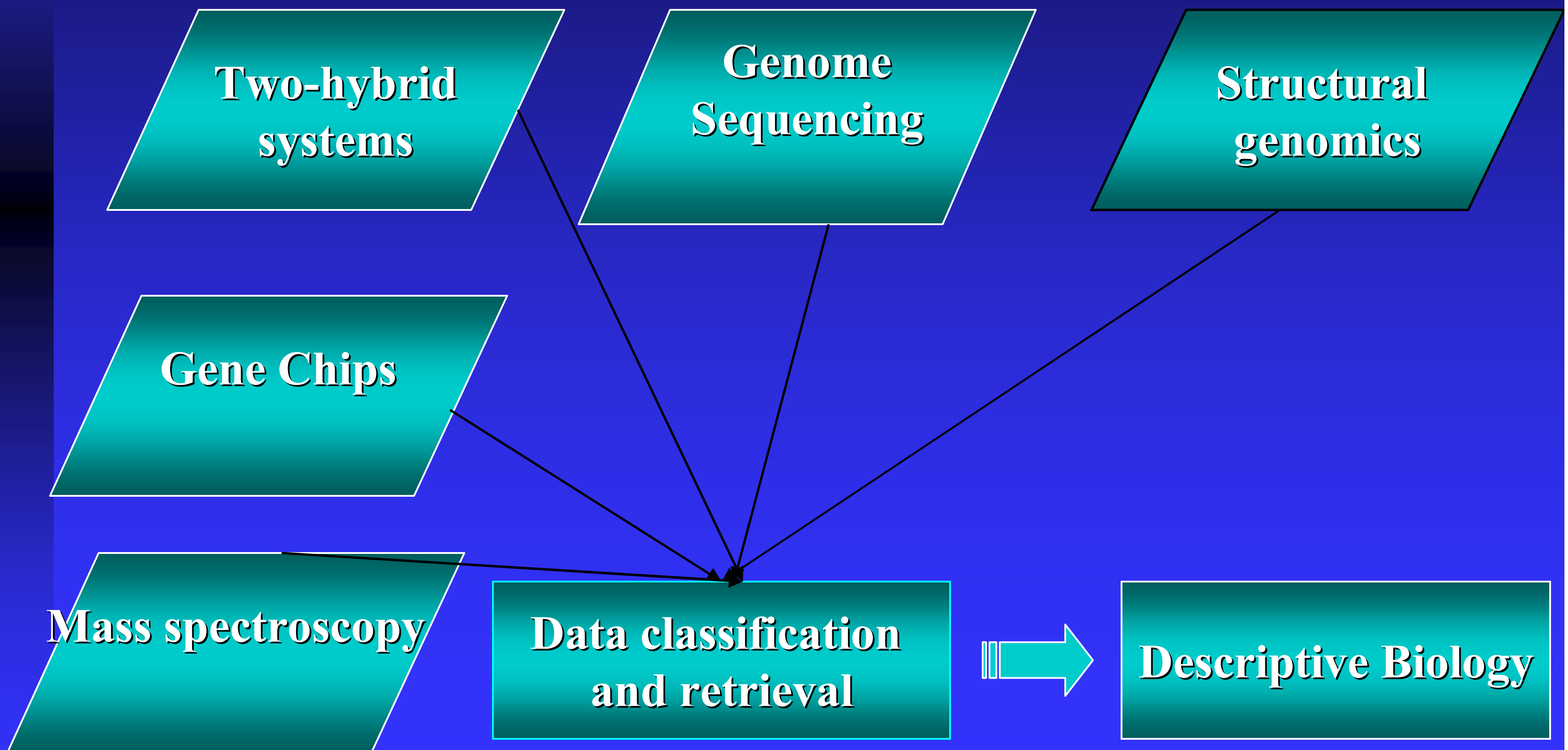
Microarray (Genoma)



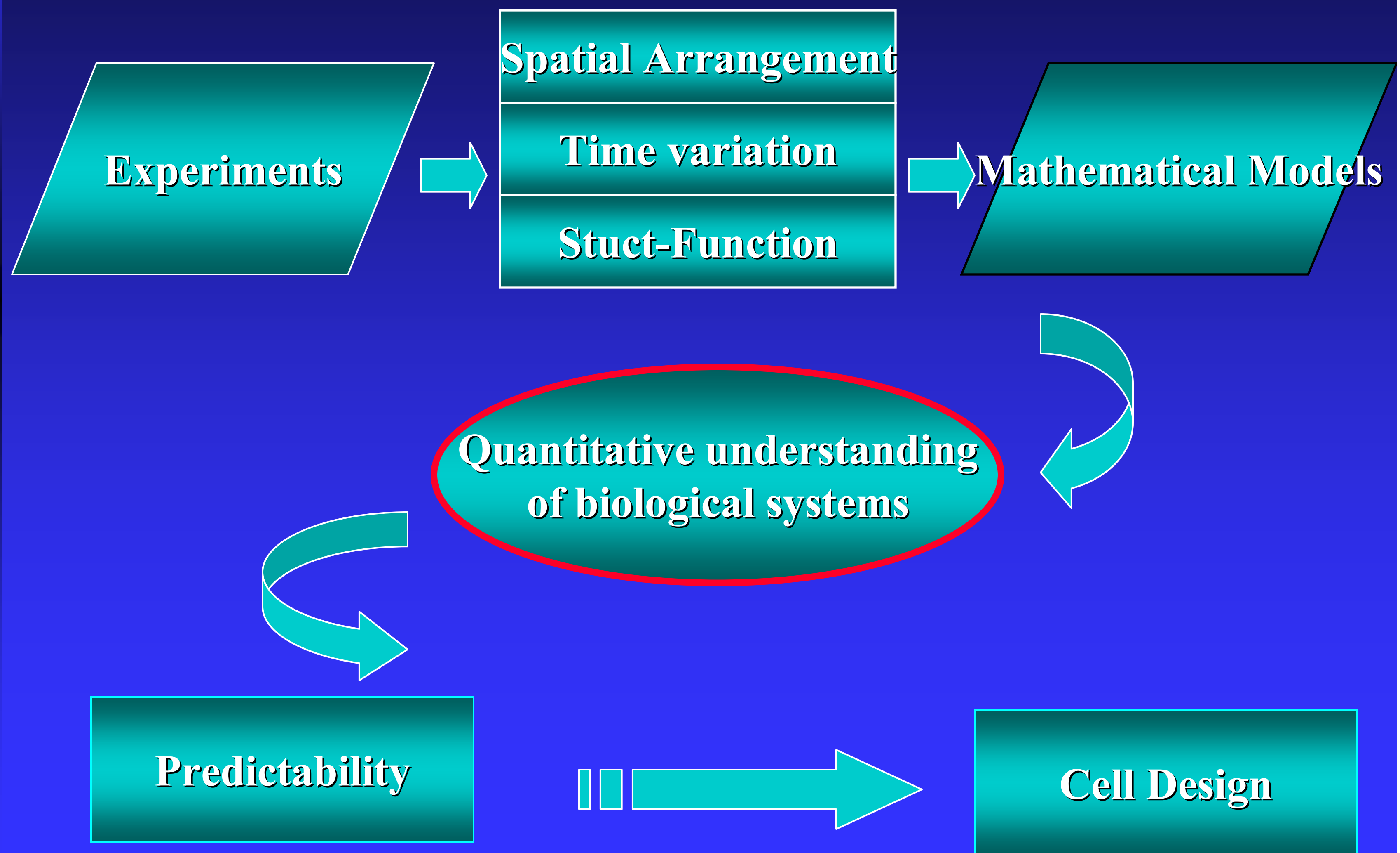
Pharmacokinetics (Metabolome)

Metabolic Pathways

Descriptive Biology



Quantitative Biology



The Dynamics of Cell Cycle Regulation

- The triumph of molecular biology of the last half of the twentieth century was to identify and characterize the molecular components of the cell, epitomized by the complete sequencing of the human genome.
- The grand challenge of <u>postgenomic</u> cell biology is to assemble these pieces into a working model of a living, responding, reproducing cell; a model that gives a reliable account of how the physiological properties of a cell derive from its underlying molecular machinery.

John J. Tyson,1* Attila Csikasz-Nagy,2,3 and Bela Novak.

Objectives

1-Understanding very simple gene networks involved in cell cycle

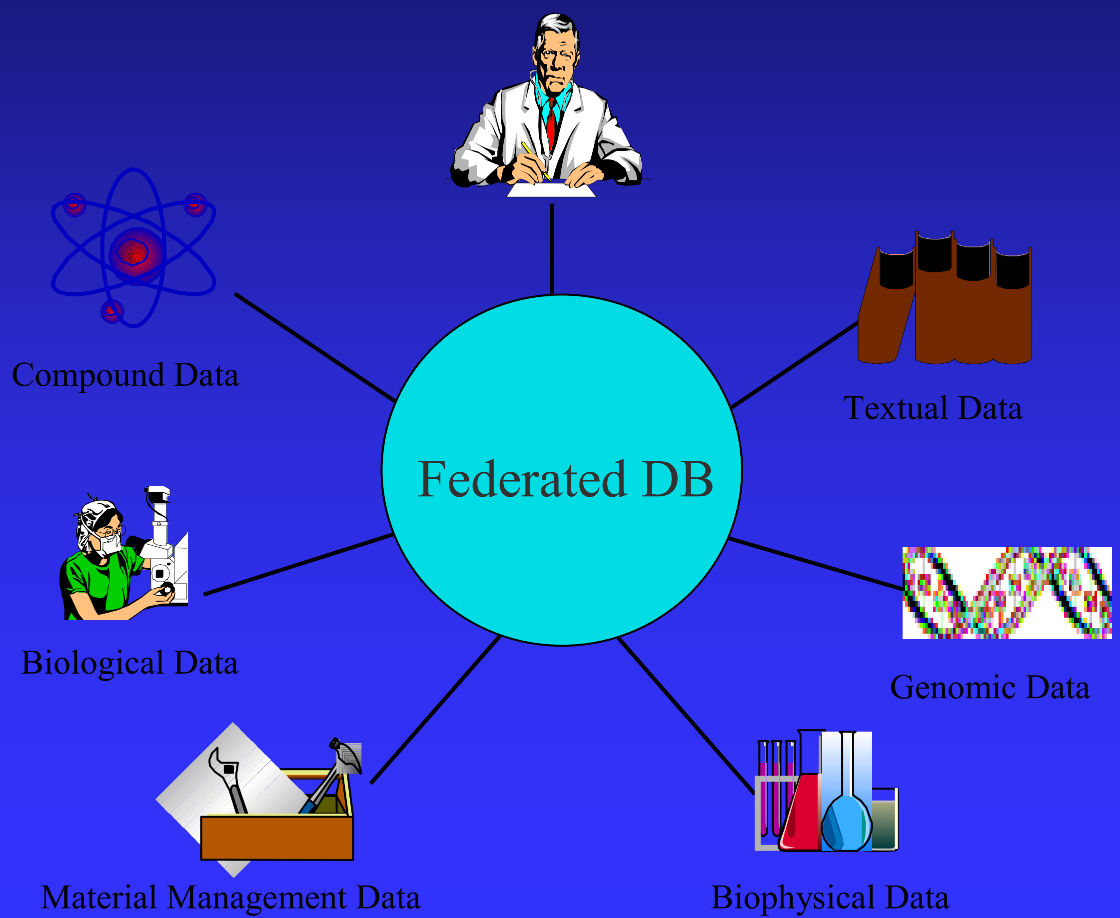
2-Obtaining general principles from those networks

3-Designing new gene networks or modify existing ones and experimental characterization.

4-Development of a computer algorithm to simulate biological processes.

5- Authomatization of gene network design

Heterogeneous Data

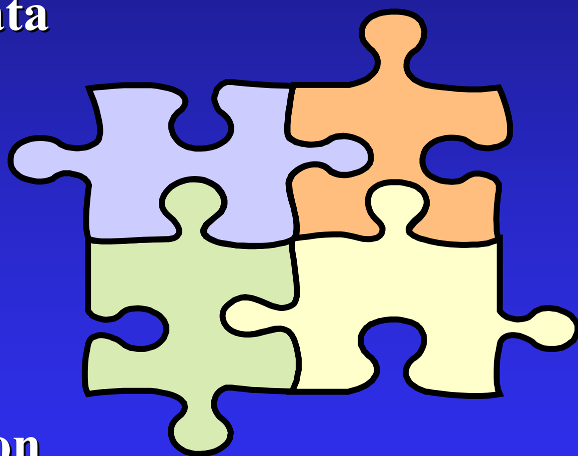


Heterogeneous Data

Textual Data

Compound Data

Proteomic Data



Toxicology Data

Genomic Data

Gene Expression Data

Clinical Data

Other Data

Heterogeneous Types of Data

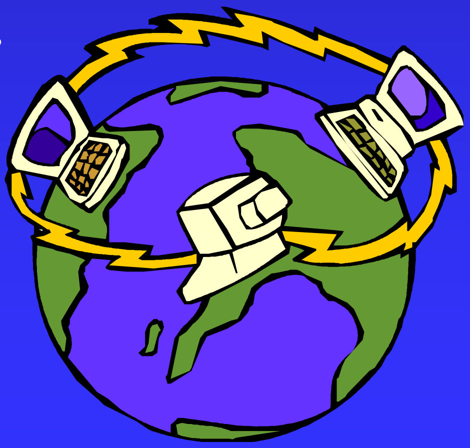
- Data sources may include:
- Relational databases
- Flat file databases
- Web pages
- Specialized search engines
- Databases may have differing data models and formats, and reside on differing platforms and operating systems

Definition of Wrappers

- Allow connection to heterogeneous sets of databases. The wrappers allow for inclusion of sources and specialized searching capabilities that can be left where they are, in their native form.
- When not available, can be written using standard software tools.
- Are currently in development for several key life science databases.
- Advantages:
- Wrappers can evolve over time.
- New wrappers can be added at any time.

Application of GRID to Biomedical Technologies

- These methodologies have high computational requirements.
- The system will be capable of using a computational architecture specifically designed for intensive computing by means of a large distributed datasets on a GRID.



Application of GRID to Biomedical Technologies

These methodologies have high computational requirements.

The system will be capable of using a computational architecture specifically designed for intensive computing by means of a large distributed datasets on a GRID.

Data Mining - text searching

- Using computers to extract information that is hidden within a large set of data.
- Data mining researchers who work with text use statistical, linguistic, and artificial intelligence techniques to go beyond simple text searching.
- **BioNLP.org** - natural language processing of biological text.
  www.ccs.neu.edu/home/futrelle/bionlp
- **BioMed Central's data mining information page**
  www.biomedcentral.com/info/about/datamining

Semantic Web

- The vision of Tim Berners-Lee for an improved version of the World Wide Web in which content is annotated in a machine readable way, to allow its meaning to be analysed by automated 'agents'.
- www.semanticweb.org
- www.mindswap.org/Science

XML

- eXtensible Markup Language (XML) is a standard text format that allows information to be represented in a structured way, thereby facilitating automatic processing.
- Different dialects of XML (known as schemas) are used to describe different types of content (for example, CML describes chemical structures, MathML describes equations).
  www.w3.org/XML
  www.xml-cml.org
  www.w3.org/Math

Ontology

- In artificial intelligence research, an Ontology refers to a structured collection of concepts relevant to a particular domain of knowledge.
- For example, an Ontology might incorporate the concept of ***engrailed***, which would be an instance of the concept **gene**.
- And like all genes, engrailed would be associated with a specific **organism** (*Drosophila*), and **chromosome** .
  Gene Ontology Consortium
- **Databases**: GENE ONTOLOGY.

Interaction Databases

- http://bind.ca/index.phtml?page=databases

- Aminoacyl-tRNA Synthetases database
- ASEdb - Alanine scanning Energetics database
- BBID - biological biochemical image database
- BIND - Biomolecular interaction network database
- BindingDB - the binding database
- Biocarta
- Biocatalysis/biodegradation database
- BioCyc knowledge library
- BioPathways consortium
- Brenda
- BRITE - Biomolecular relations in information transmission and expression
- COMPEL (composite regulatory elements)

Interaction Databases

- COPE - cytokines online pathfinder encyclopaedia
- CSNDB - cell Signaling networks database / CSNDB paper
- Curagen Pathcalling
- DIP - database of interacting proteins
- DPInteract - DNA-protein interactions
- DRC - database of ribosomal Crosslinks
- Dynamic Signaling maps
- EMP - the Enzymology database
- ENZYME - enzyme nomenclature database
- FIMM - A database of functional molecular immunology
- FlyNets - gene networks in the fruit fly

Interaction Databases

- GeneNet (Gene networks)
- GeNet - Gene Networks Database
- HIV Molecular Immunology Database
- HOX Pro db - Homeobox Genes DataBase
- InBase - The Intein Database
- Indigo (Gene networks)
- Interact - A Protein-Protein Interaction database
- Inter-Chain Beta-Sheets (ICBS) - A database of protein-protein interactions mediated by interchain beta-sheet formation
- JenPep: Immunology MHC-peptide database
- KEGG - Kyoto Encyclopedia of Genes and Genomes
- Kohn Molecular Interaction Maps

Interaction Databases

- MDB - Metalloprotein Database and Browser
- MHCPEP - A database of MHC binding peptides
- MINT - a database of Molecular INTeractions
- MIPS Yeast Genome Database
- MMDB - Molecular Modeling Database
- NetBiochem Welcome Page
- ooTFD - object-oriented Transcription Factors Database)
- ORDB - Olfactory Receptor Database
- PATIKA - Pathway Analysis Tool for Integration and Knowledge Acquisition
- PFBP - Protein Function and Biochemical Pathways Project
- PhosphoBase - A database of phosphorylation sites
- PIM (Protein Interaction Map)
- PIMdb - Drosophila Protein Interaction Map database

Interaction Databases

- PKR - Protein Kinase Resource
- ProChart Database (at AxCell Biosciences)
- ProNet Online - Protein Interactions on the Web (Myriad)
- REBASE - The Restriction Enzyme Database
- Relibase - A program for searching protein-ligand databases
- RegulonDB - A DataBase On Transcriptional Regulation in E. Coli
- SELEX_DB
- SoyBase
- SPAD - Signaling Pathway Database
- SPIN-PP - Surface Properties of INterfaces - Protein Protein Interfaces
- STKE - Signal Transduction Knowledge Environment
- SYFPEITHI - A Database of MHC Ligands and Peptide Motifs

Interaction Databases

- TRANSFAC - The Transcription Factor Database
- TRANSPATH - Signal Transduction Browser
- TRRD - Transcription Regulatory Regions Database
- WIT (What Is There?) - Interactive Metabolic Reconstruction on the WEB
- **Small Molecule Databases**:
- The Amino Acids (at Freie Universita:t Berlin)
- CSD - The Cambridge Structural Database
- ChemIDPlus
- Klotho - Biochemical Compounds Declarative Database
- LIGAND - Database for enzymes, compounds, and reactions
- LIPIDAT
- MathMol - Mathematics and Molecules (Molecular Library)
- Molecular Models from Chemistry at Okanagan University College