


Slide 1 - Slide 1

The slide features a dark blue background with a faint grid pattern. A central graphic consists of overlapping, wavy lines in shades of purple, blue, and orange, resembling a DNA double helix or a data visualization. The text "Bioinformatics Overview" is displayed in a light blue, sans-serif font. Below it, the text "VIBE Education Edition (VIBE-Ed) Initiative" is written in a white, italicized serif font. In the bottom left corner, there is a logo consisting of a blue and white 3D cube-like structure. In the bottom right corner, the word "INCOGEN" is written in a light blue, spaced-out, sans-serif font.

Bioinformatics Overview

*VIBE Education Edition (VIBE-Ed)
Initiative*



INCOGEN

Slide notes

This presentation will give you a brief overview of bioinformatics,

Outline

- Bio-"technology" - Then and Now
- Bioinformatics - Golden Age?
- Challenges in Bioinformatics
- VIBE

INCOGEN

Slide notes

both its history and its present state. Included in this overview is a discussion of whether or not we have reached the "Golden Age" of bioinformatics. We will then begin a discussion of the short- and long-term challenges that bioinformatics faces, and will conclude with a discussion of how VIBE can help to meet these challenges.

Origins of Biotechnology

Early Speculation

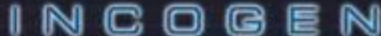

- 6000 BC: Yeast was used to make beer by Sumerians and Babylonians.
- 400 BC: Hippocrates - "male contribution to a child's heredity is carried in the semen"
- 320 BC: Aristotle - "all inheritance comes from the father. Female babies are caused by 'interference' from the mother's blood."
- 100 AD: Romans speculated that mares can be fertilized by the wind.
- 1673 AD: Anton van Leeuwenhoek describes protozoa and bacteria; confirms existence of sperm cells.

Things are getting smaller

- 1859: Charles Darwin - "*On the Origin of Species*"
- 1865: Gregor Mendel - laws of heredity, internal units of information: later become known as *genes*

Converging on DNA

- 1900: The science of genetics is born.
- 1953: Watson and Crick propose the double-stranded, helical, complementary, anti-parallel model for DNA.

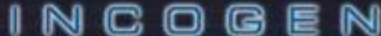



Slide notes

First, a little about the history of biotechnology and bioinformatics: Deliberate uses of biotechnology have existed for more than 8000 years. Here we see a brief timeline of how biotechnology has evolved over that time period.

Origins of Biotechnology, cont'd The Dawn of "Modern" Biotech...

- **1977: Genentech** reports the production of somatostatin.
- **1980: USSC** rules in that genetically altered life forms can be patented.
- **1981:** Scientists at Ohio University produce a transgenic mouse.
- **1985:** Genetic fingerprinting enters the court room.
- **1989:** Creation of NCHGR
- **1990:**
 - Launch of Human Genome Project (Projected duration: 15 years, projected cost: \$13 billion)
 - First gene therapy - ethics concerns
 - Publication of "Jurassic Park"
- **1994:**
 - **Flavr Savr** gains FDA approval
 - The first crude but thorough linkage map of the human genome appears.
- **1995:**
 - DUMC researchers transplant hearts from genetically altered pigs into baboons
 - *H. influenzae* sequence completed
- **1998:**
 - *C. elegans* is sequenced
 - A rough draft of the human genome map is produced
 - Celera is founded - Venter proposes new method to sequence the genome
 - And... INCOGEN is founded!

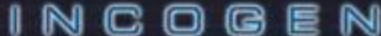



Slide notes

Genentech's production of somatostatin in 1977 marked the beginning of what we consider "modern" biotechnology and opened a host of legal and ethical issues. Here we see a brief timeline of major events in modern biotech.

Biotechnology Now: Golden Age ... ?

- Human genome sequence completed
- Tens of model organism sequenced
- Hundreds of new biotechnology companies
- Integral component of pharmaceutical and agricultural research
 - Pharmaceutical benefits
 - Agricultural research benefits
- Key source of economic growth and employment



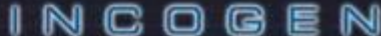

Slide notes

It appears at first glance that the biotechnology industry is flourishing and producing important and necessary results. The human genome and hundreds of organisms have been completely sequenced. There are hundreds of new biotechnology companies, and genetics has become an integral part of pharmaceutical and agricultural research. Because of biotechnology and genetics research, we have improved medical diagnoses; treatments for cancer, hepatitis, diabetes, HIV/AIDS, malaria, asthma, arthritis, Alzheimer's, and cardiovascular disease and drought and pathogen resistant crops with improved nutritional value and higher yields. Biotechnology is a key source of economic growth and employment.

Biotechnology Now: or Hype ... ?

- Billions of dollars spent on research that didn't go anywhere
- Unfulfilled promises, conflicting egos, lack of standards and interoperability, IP conflicts...
- "The VC bubble", dozens of companies floundered
- Genomics doesn't yield all the answers...
- How many genes are there, finally?

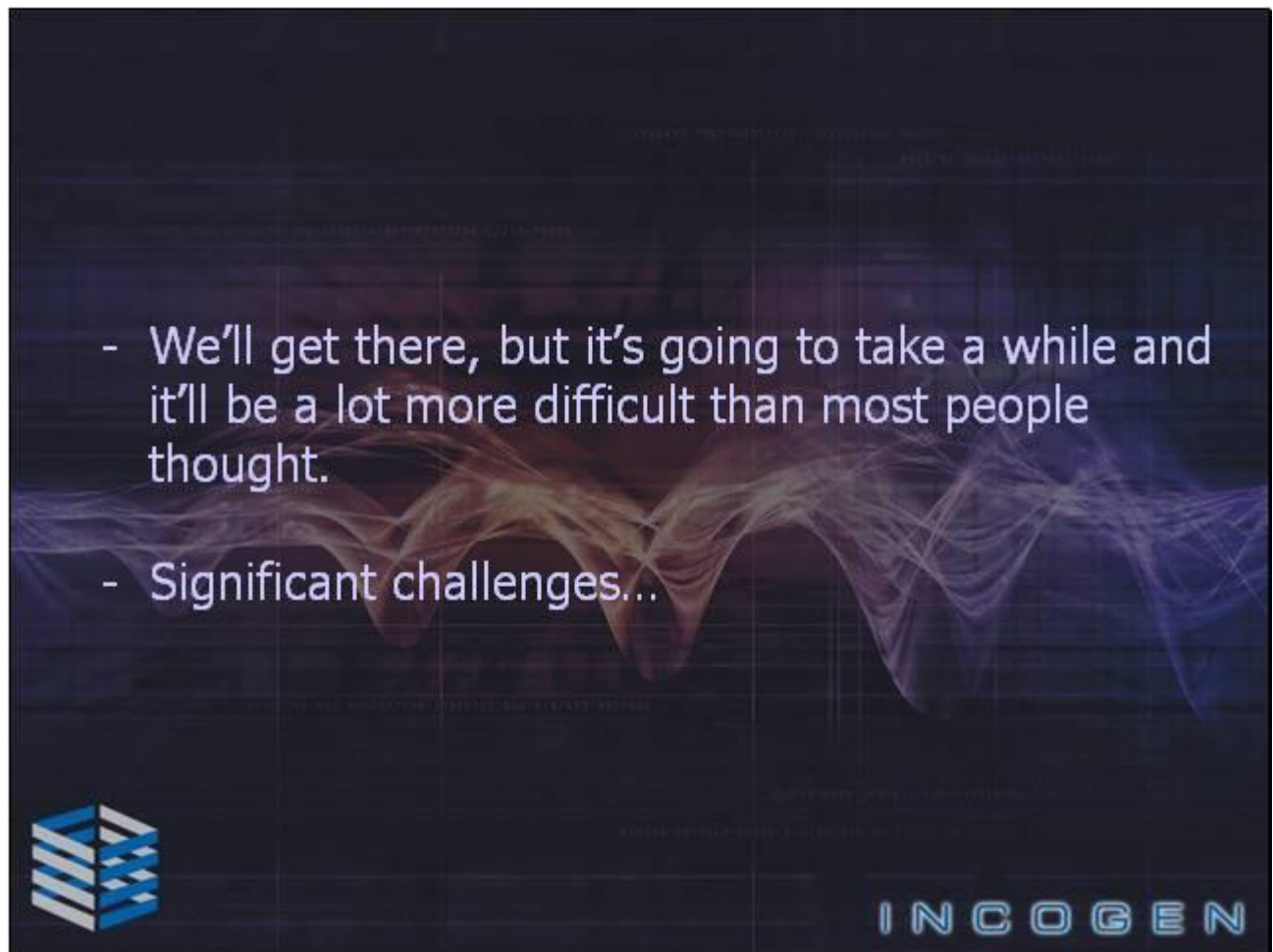
So which is it?



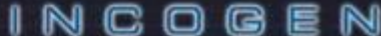

Slide notes

However, billions of dollars have been spent on biotechnology research that didn't always lead to the answers. There is a lack of standards, which makes it difficult to share data; conflicting egos have caused work to be duplicated. Dozens of biotechnology companies floundered when the "Venture Capital bubble" burst. We don't even know how many genes exist.

Slide 7 - Slide 7

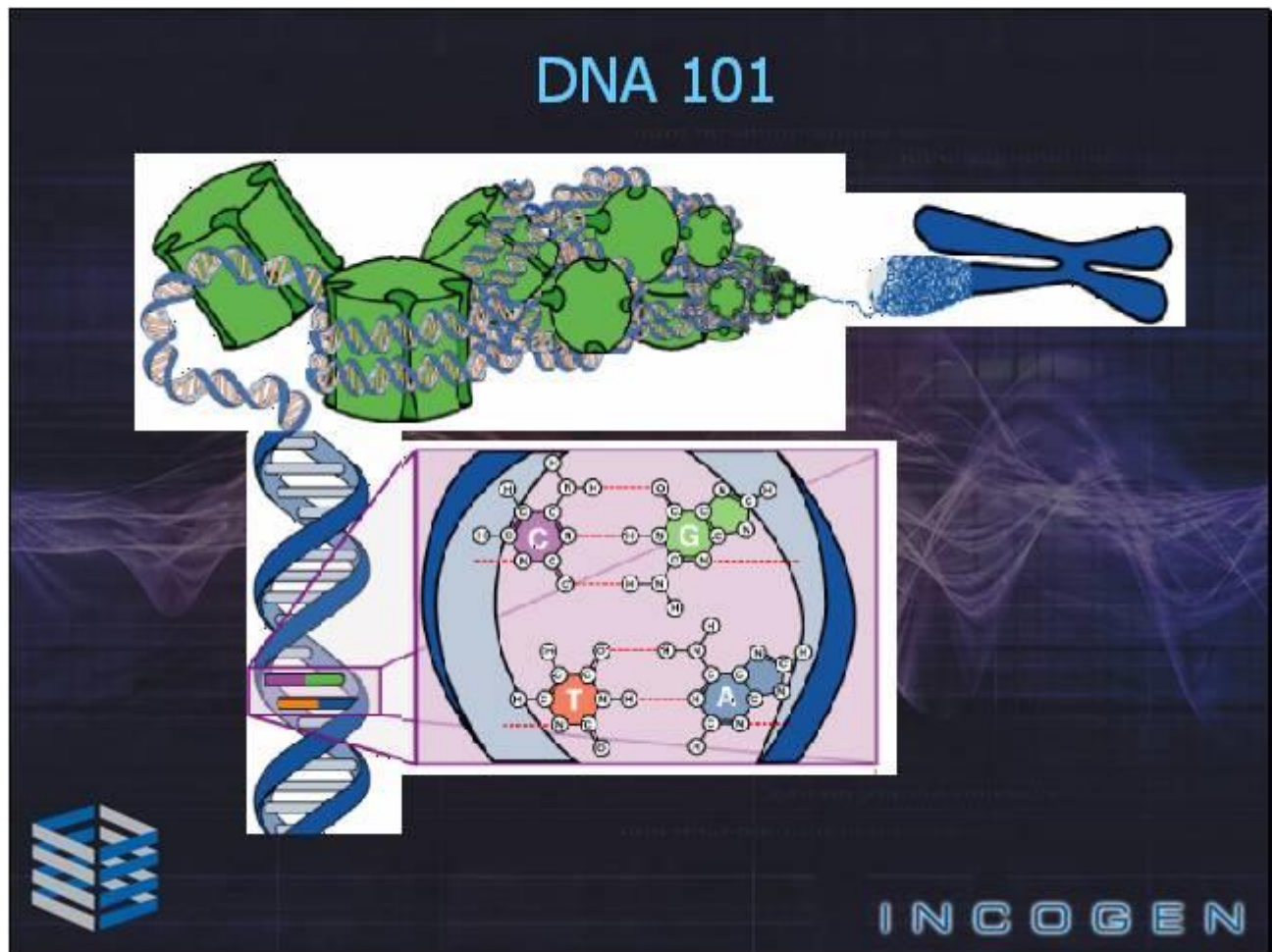
The slide features a dark blue background with a faint grid and abstract, glowing orange and purple wave-like patterns. The text is centered in a white, sans-serif font. In the bottom left corner, there is a logo consisting of a blue and white 3D cube-like structure. In the bottom right corner, the word "INCOGEN" is written in a blue, stylized, outlined font.

- We'll get there, but it's going to take a while and it'll be a lot more difficult than most people thought.
- Significant challenges...



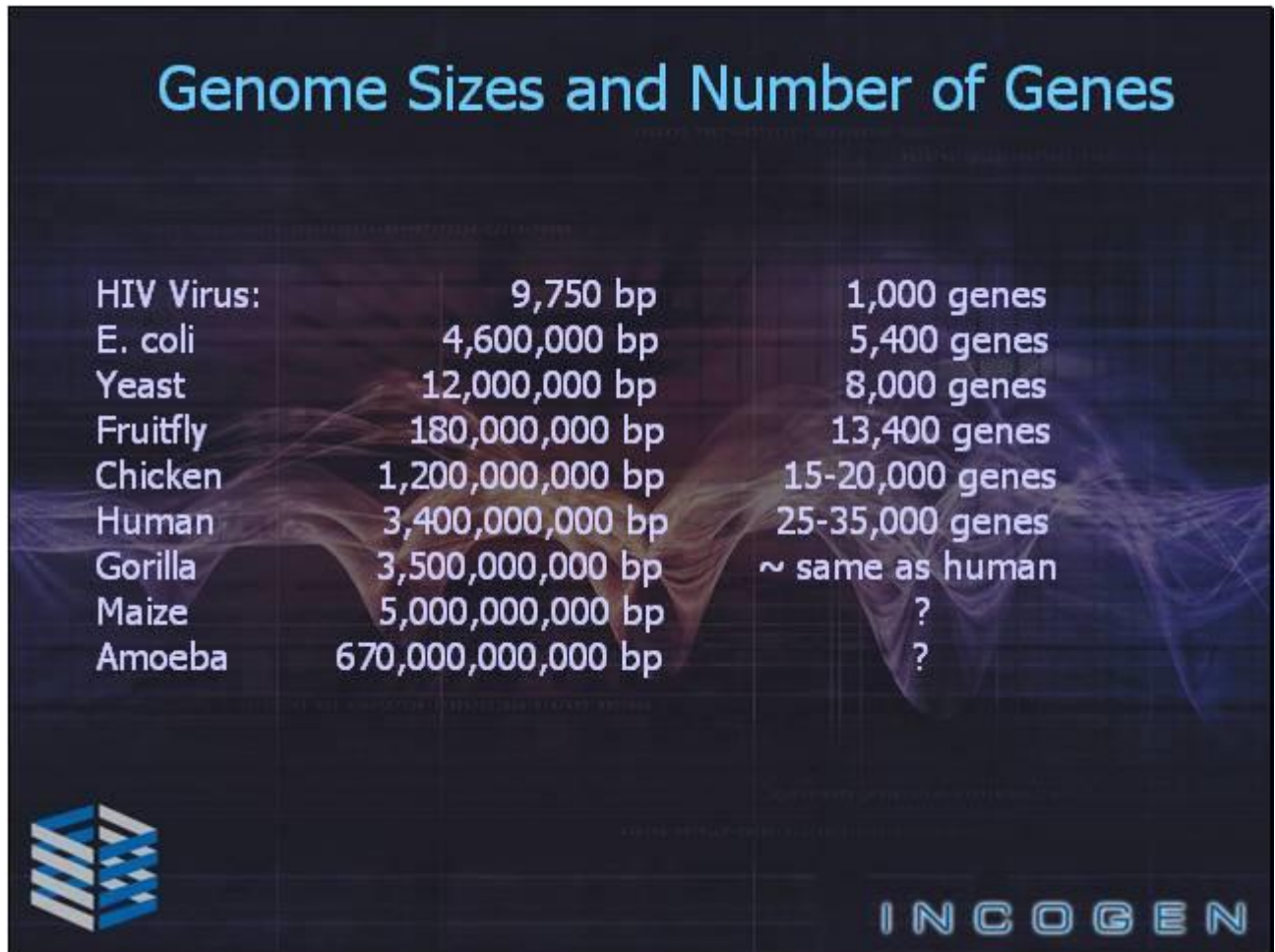
Slide notes

We have made a lot of progress in the last century, and progress will continue in the future. Compared to disciplines such as physics and chemistry, biotechnology is in its infancy. It's going to take a lot more effort and will be more difficult than we initially thought. Some significant challenges await us.



Slide notes

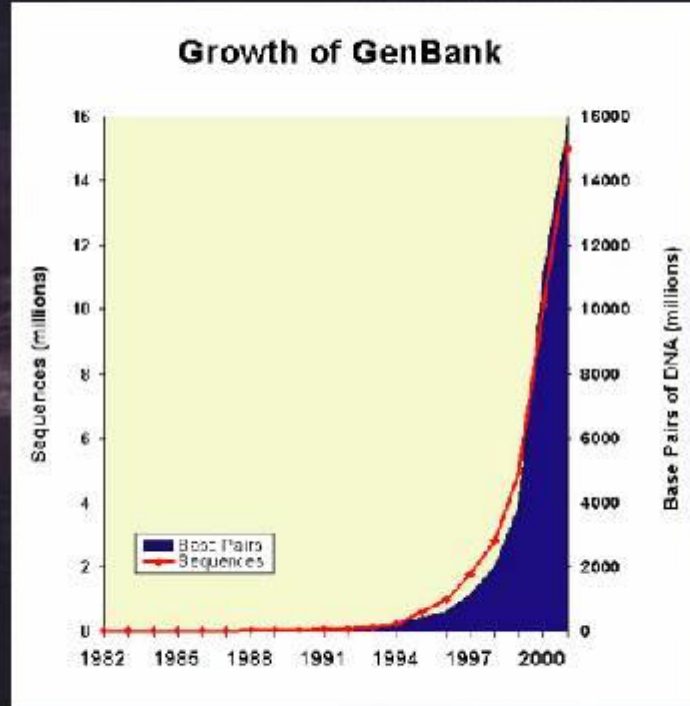
Before we discuss the challenges we have to meet, I want to give a brief overview of DNA and genetics. A chromosome is composed of 50 million to 250 million base pairs and is approximately 5 microns long. The strands are held together with hydrogen bonds between Cytosine and Guanine residues and between Thymine and Adenine residues. Each base has a vertical rise of 0.34nm, and the double helix makes a complete turn every 3.4nm of its length. Humans have approximately 3.4 billion base pairs in their chromosomes and two sets of chromosomes per cell, so the total amount of DNA in each human cell would be 1-2m long if stretched out.



Slide notes

The number of base pairs and the number of genes that are associated with an organism varies widely. Not surprisingly, viruses tend to have the smallest genomes and the fewest genes. More complicated organisms tend to have larger genomes and more genes, so it may come as a surprise to find that an amoeba's genome is almost 200 times longer than a human's. Exons, or regions of the genome that code for proteins, tend to range in length from a few base pairs to a few hundred base pairs, while the introns, or regions of the genome that do not code for proteins, range in length from tens of base pairs to a few thousand base pairs.

Challenge #1: Data Explosion



INCOGEN

Slide notes

The first challenge for biotechnology, and one that only promises to get worse as time progresses, is data explosion. The graph you're looking at represents the number of sequences and base pairs stored in GenBank over a twenty year period. In 1982, a student could earn their Ph.D. for sequencing a single gene. Today, Celera and companies like it can sequence 100 million bases a day. The internet and repositories like GenBank allowed major sequencing efforts to succeed; the sequencing technology was developed in parallel and results from one organization could quickly and easily be used by others in their efforts. However, the challenge is to continue developing methods to quickly and efficiently search through existing data as that data repository grows exponentially.

Challenge #2: Data Heterogeneity


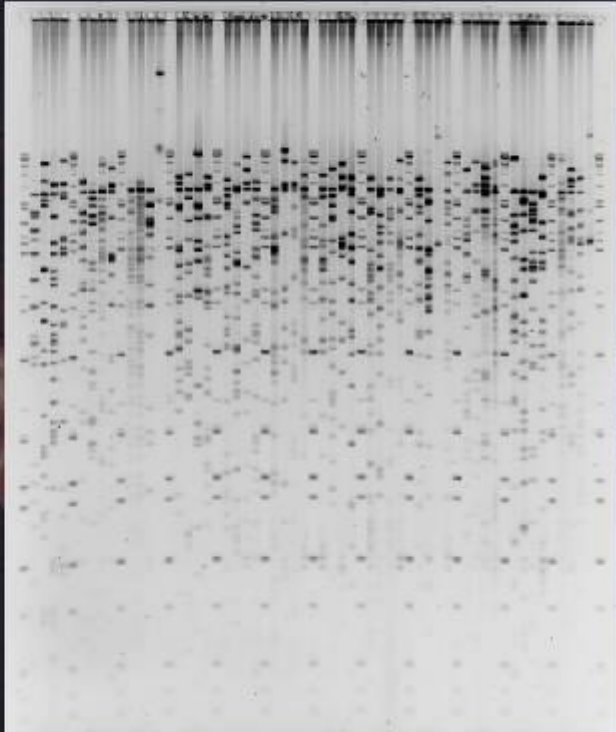
The slide features a central collage of various data visualization types. At the top center is a large, colorful network diagram with numerous nodes and connecting lines. To its left is a vertical heatmap with many small colored squares. Below the heatmap is a circular inset containing a cartoon illustration of three people sitting at a desk with a computer monitor. To the right of the cartoon is a DNA microarray image showing multiple lanes of colored spots. Below the microarray is a chromatogram with several peaks. At the bottom left is a 3D protein structure model. At the bottom right is a gel electrophoresis image with multiple lanes of bands. The word 'INCOGEN' is written in a stylized, glowing blue font at the bottom right of the slide.

Slide notes

Another big challenge is data heterogeneity. Currently, there are multiple methods for getting information from DNA and proteins, and the disparate data types must be stored using a common format if it is to be accessible by everyone.

DNA Fingerprinting

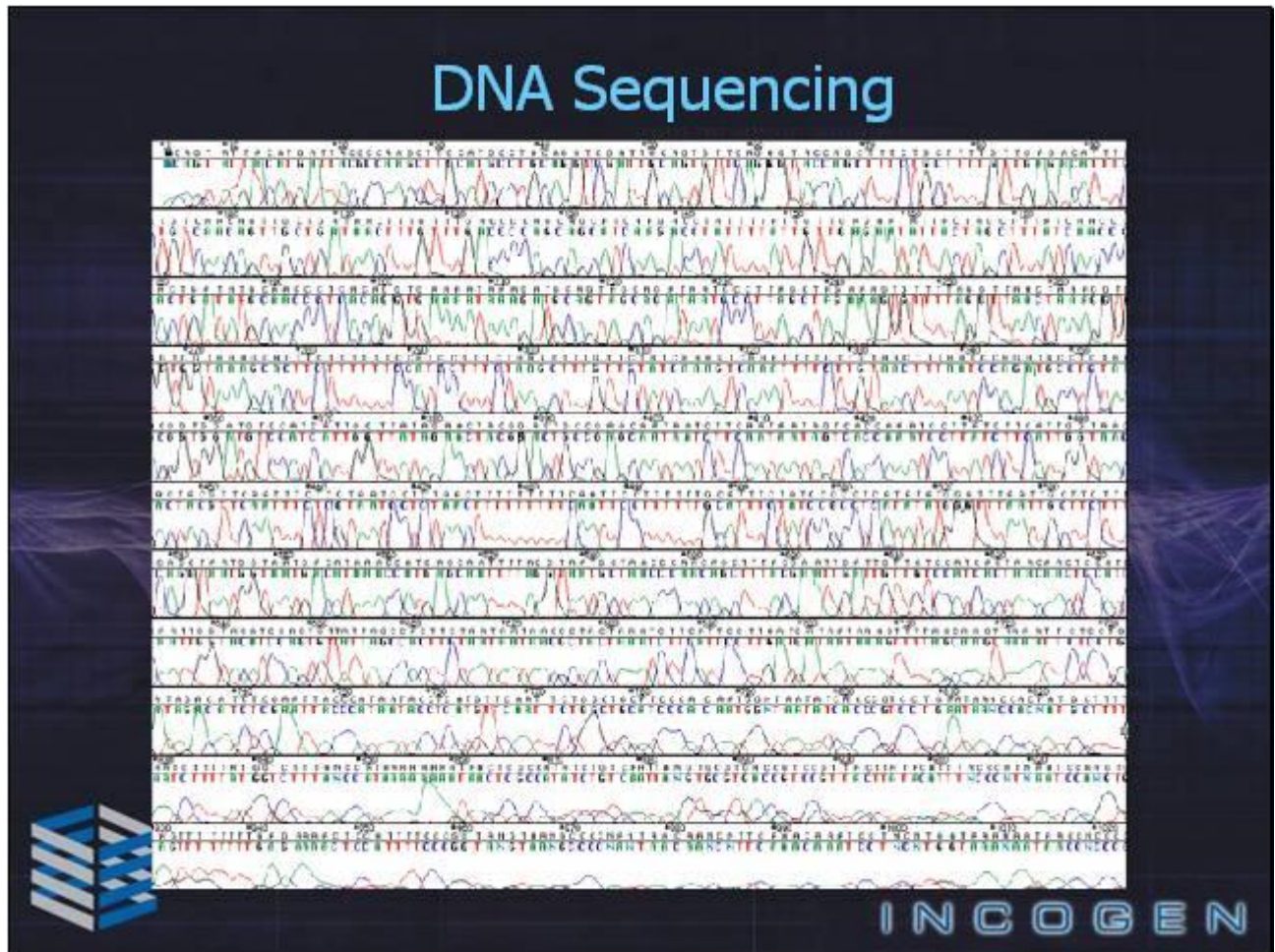
- Gel electrophoresis
- separation by charge/mass ratio through electrophoresis
- unique identity of each clone based on its DNA sequence restriction pattern



INCOGEN

Slide notes

For example, this image is raw data from a DNA fingerprinting experiment. DNA fingerprinting uses gel electrophoresis to separate pieces of DNA based on charge/mass ratio. The DNA to be analyzed is broken apart before the electrophoresis using a known restriction enzyme. When the DNA pieces are run through the gel, the pattern formed is unique to the clone being analyzed. The output of the experiment is an image, such as this one, that shows the final positions of the DNA segments.

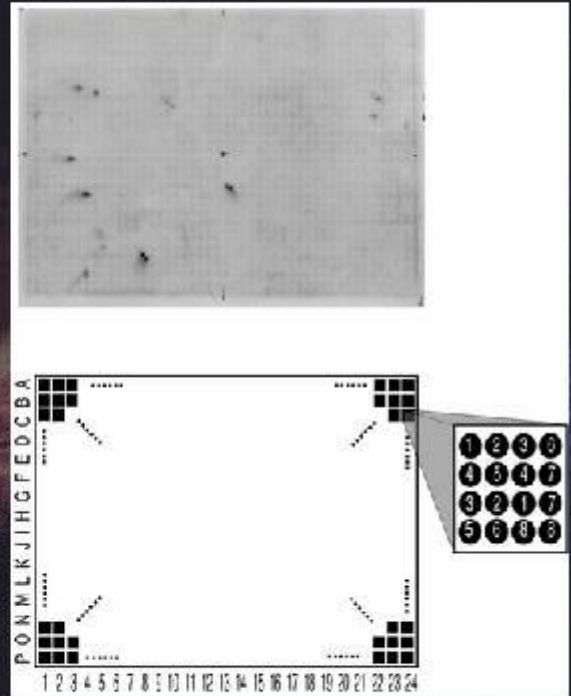


Slide notes

DNA sequencing leads to a different output type. Segments of DNA are split into two strands, the strands are bound to a substrate and then treated with fluorescently-labeled A, G, C, and T. The resulting substrate is run through a machine that can detect the colors associated with the labeled bases. The results appear as a graph of the intensity of each color with position in the sequence and the analysis program's best guess of which residue is present at each position.

DNA Macroarrays

- association of two **complementary** nucleic acid strands
- strand acting as the probe is radiolabeled
- hybridization event ("hit") occurs when the probe hybridizes to a clone containing the complementary sequence
- "hits" can be scored and a matrix of hits vs. probes can be created




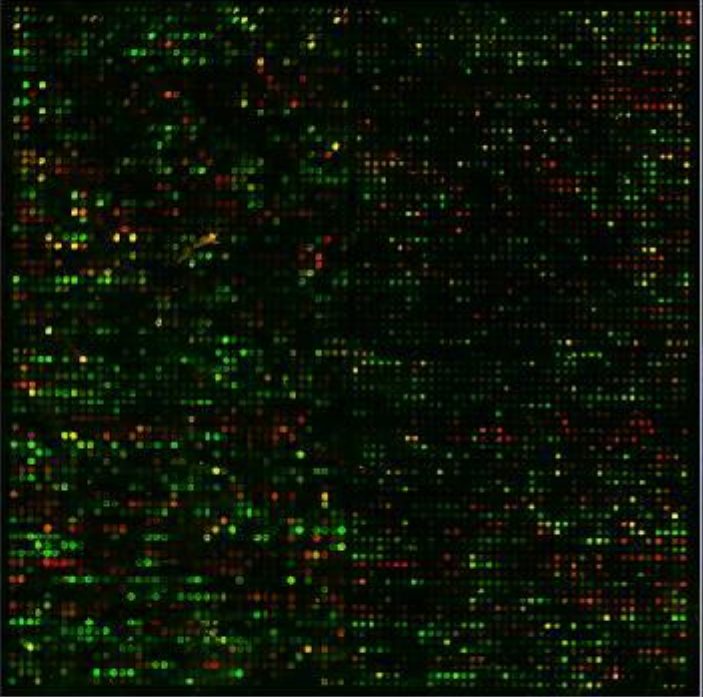
INCOGEN

Slide notes

DNA macroarray experiments help scientists determine when genes are expressed. A known set of genes are attached to a macroarray slide. DNA from a cell at a particular experimental condition are washed over the slide and those genes which were expressed will "hybridize" with their complement on the slide. The darker the spot, the more concentrated the gene.

DNA Macroarrays

- Array sequences corresponding to genes onto glass slides
- Study expression patterns of those genes

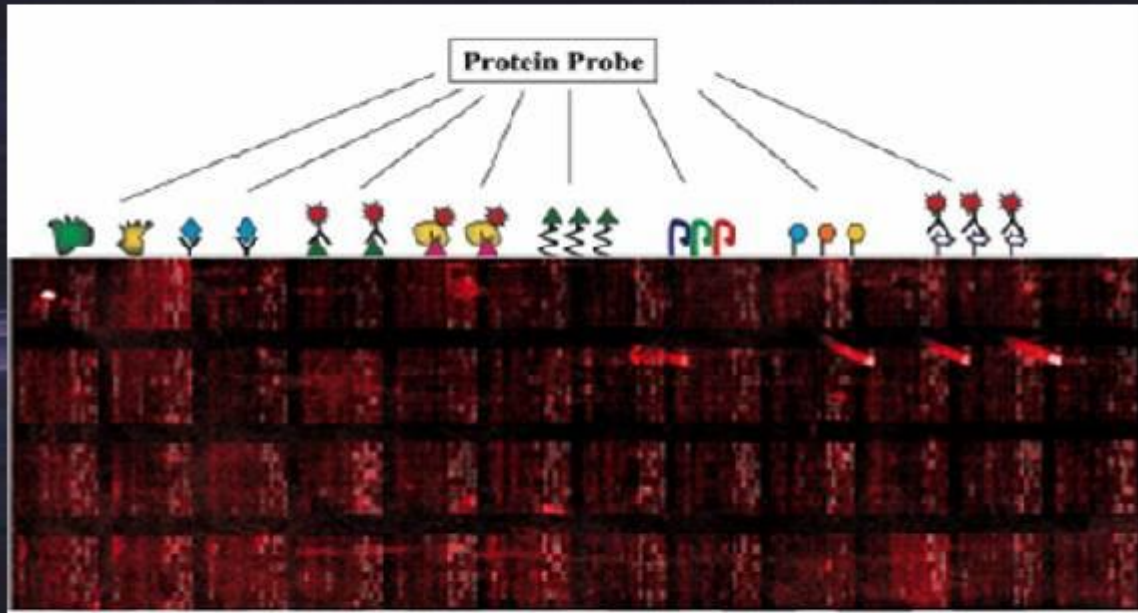


INCOGEN

Slide notes

DNA microarray experiments produce yet another output type. The experiments themselves are very similar to DNA macroarray experiments. In this case, the sequences are put on a glass slide, sometimes at densities of 20,000 spots (or genes) per slide. The DNA that is washed over the slide is radio-labeled, and the brightness of each spot indicates how much of that gene is present. An example of the output type of these experiments is shown here.

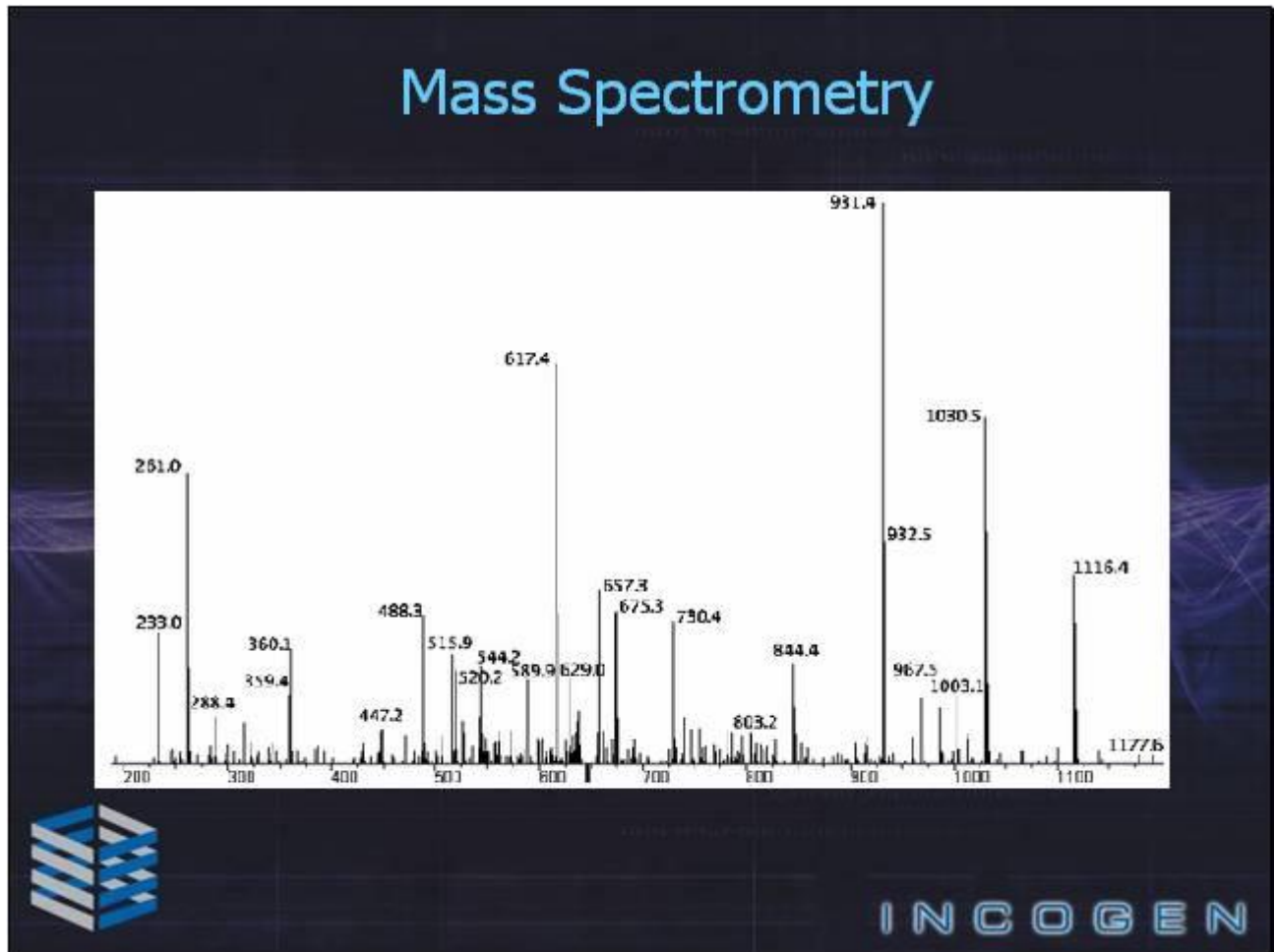
Protein/Antibody Microarrays



INCOGEN

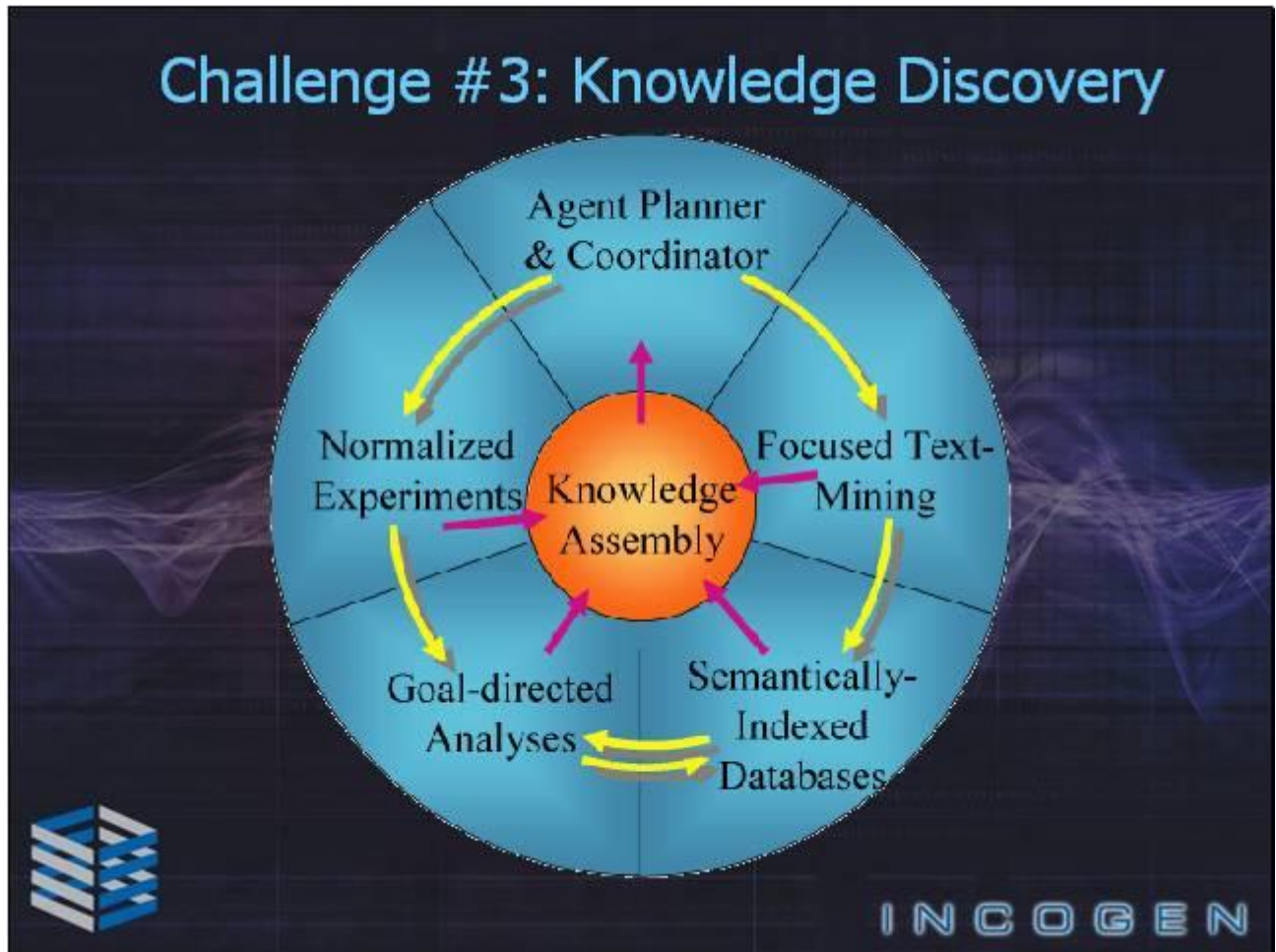
Slide notes

Protein microarrays present yet another set of data. These microarrays provide for the analysis of protein expression, including antibody response to disease.



Slide notes

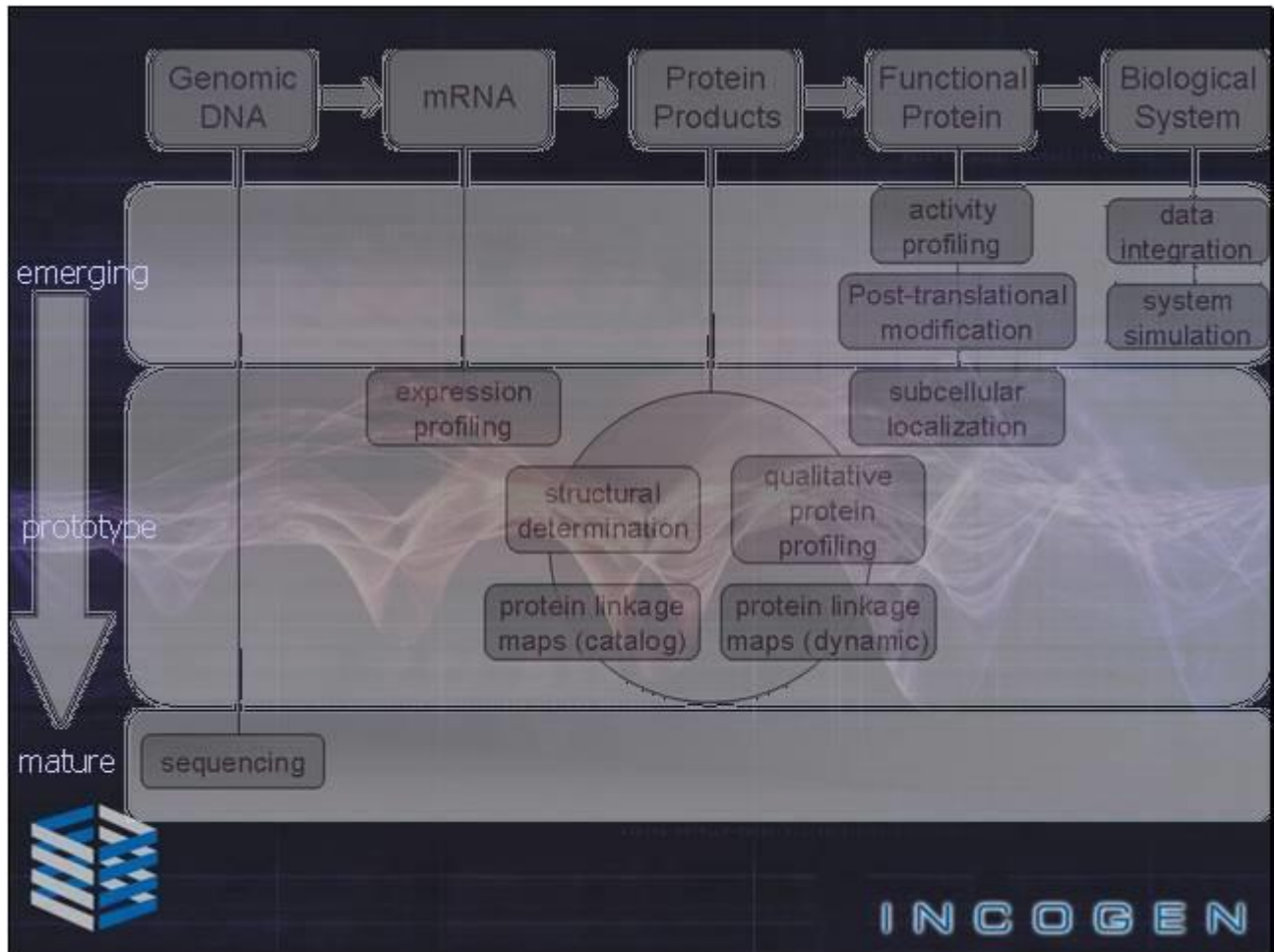
Mass spectrometry, which produces results similar to this, is frequently used to identify proteins. All of these experiments are used to identify either DNA, expressed genes, or proteins. In an ideal world, data which are related should be connected in such a way that it is easy to find all of the data related to a particular protein or expressed gene. The differences in the output from the various experiments make this difficult.



Slide notes

Solutions to the first two challenges will help to solve the third - knowledge discovery. It is a waste of time and resources to repeat experiments that have already been done solely because the information from the previous experiments is not readily accessible. Fortunately, we are working on solutions to this all of the time, and just as the internet helped the Human Genome Sequencing Project, new technologies that are currently being developed, such as the semantic web and ontologies, will make this easier.

Slide 19 - Slide 19



Slide notes

Why are there so many data types? Because we need data from all stages of the process. As we move further along, the technologies become less mature, in part because things become much more complicated and require the combination of expertise in different fields. For example, activity profiling has recently become of great interest; this involves measuring protein kinetics: where and when proteins move, how fast they move, the mechanisms of movement, degradation, and recycling. These are all important to understanding the life cycle of a gene's response to stimuli, and the study is just now getting underway.

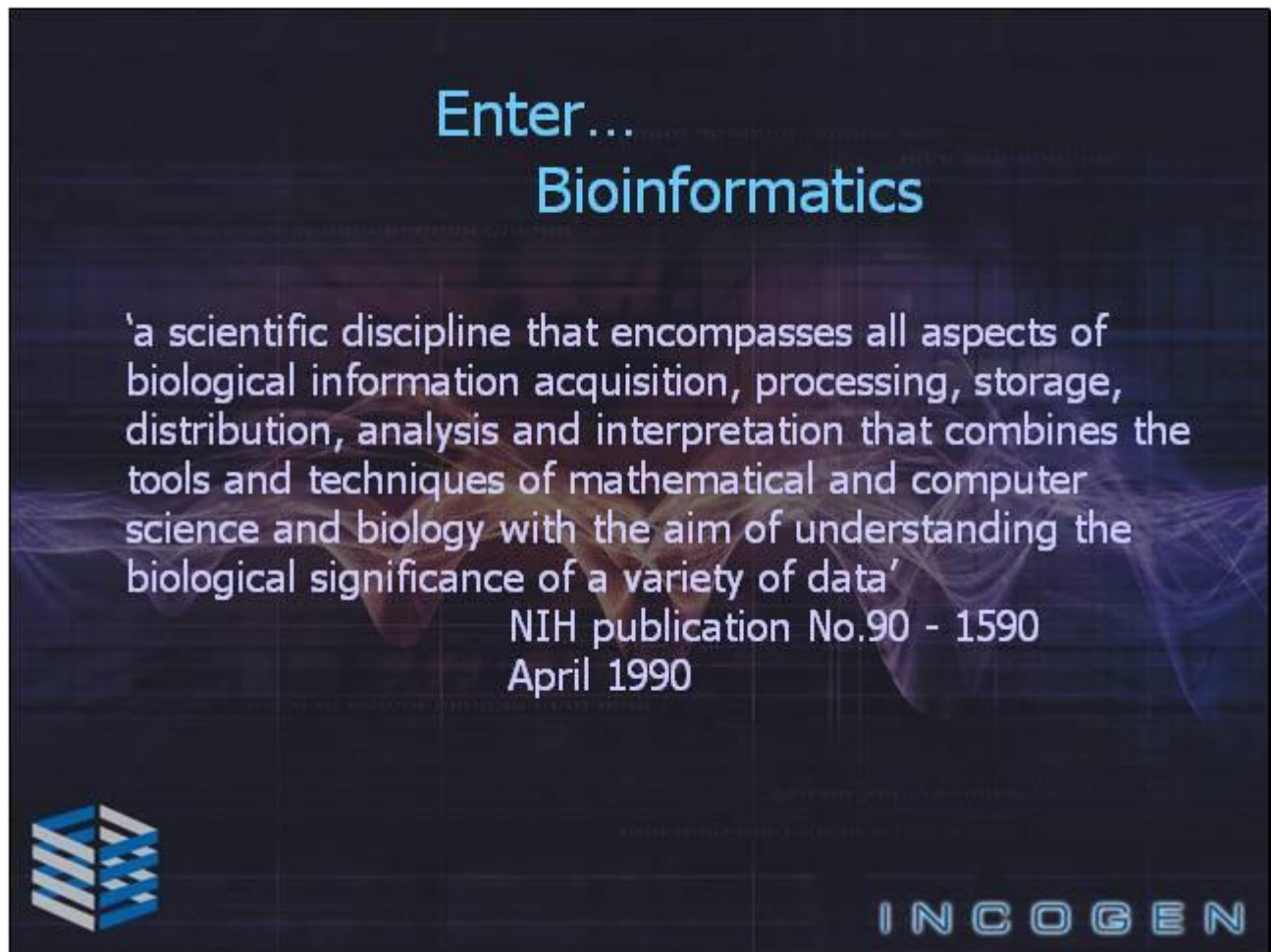
Integrative Biology/Science

- Requires the melding of traditionally divergent disciplines, such as:
 - Molecular biology
 - Clinical research
 - Physics
 - Computer Science
 - Statistics,
 - Etc.



Slide notes

The one thing that we have learned through all of this is that no single experiment can hope to answer all of the questions and no single discipline can hope to gain all of the understanding that we need to fully use and appreciate biotechnology. It will require a melding of disciplines that traditionally have worked as separate entities, such as molecular biology, chemistry, physics, computer science, and statistics. We need formal collaborations and tools designed to deal with the disparate types of data produced and needed by each discipline in order to facilitate data exchange between researchers.



Enter... Bioinformatics

'a scientific discipline that encompasses all aspects of biological information acquisition, processing, storage, distribution, analysis and interpretation that combines the tools and techniques of mathematical and computer science and biology with the aim of understanding the biological significance of a variety of data'

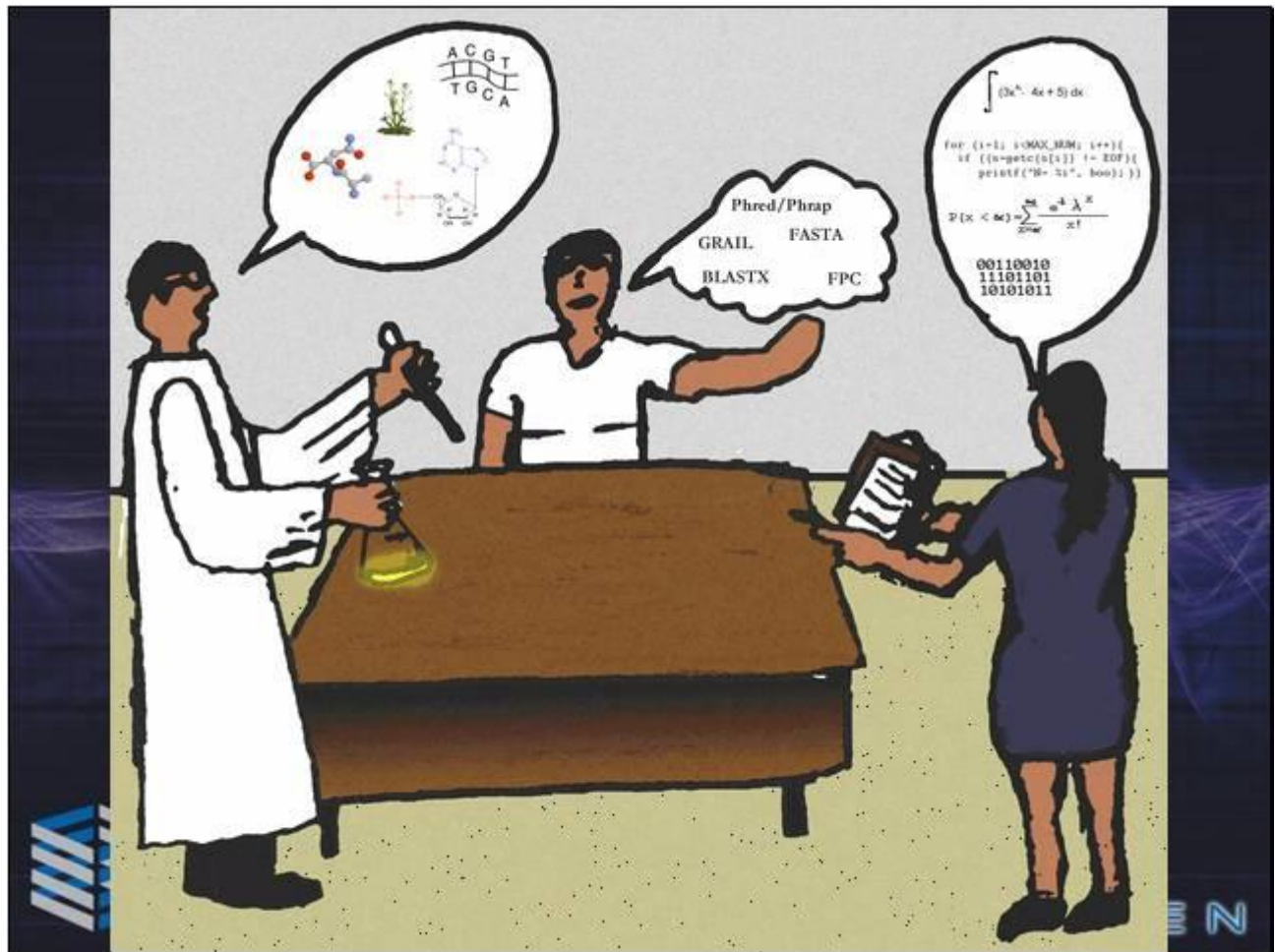
NIH publication No.90 - 1590
April 1990

INCOGEN

Slide notes

Thus, we need bioinformatics, a “scientific discipline that encompasses biological information acquisition, storage, distribution, analysis and interpretation that combines the tools and techniques of mathematical and computer science and biology with the aim of understanding the biological significance of a variety of data.”

Slide 22 - Slide 22

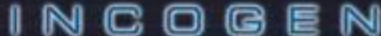



Slide notes

Bioinformatics allows the biologist with his wet lab experiments, the statistician with his algorithms, and the computer scientist with his scripts to work together.

Roles of Bioinformatics

- Provide infrastructure to conduct projects
- Analysis and presentation of data
- Research and development of new analysis tools



Slide notes

Bioinformatics supplies an infrastructure for researchers to conduct projects, allows for the analysis and presentation of data, and provides for the research and development of new analysis tools, such as VIBE.

What is VIBE?

- **Visual Integrated Bioinformatics Environment**
- **Bioinformatics Workflow Management**
 - Drag-and-drop pipeline creation
 - Analysis parameterization
- **Bioinformatics Exploration Platform (Workbench)**
 - Step-by-step process creation, filtering and visualization
 - Archives and templates
- **Technical Highlights**
 - Visual programming applied to bioinformatics
 - Distributable, grid-compatible, scalable, extensible
 - Simple interface for centralized access to tools and data sources
 - Built-in reproducibility and tracking (archives, templates, execution monitoring, etc.)



Slide notes

VIBE stands for the "Visual Integrated Bioinformatics Environment." It allows users to create and manage their own analysis pipelines. The interface is clear and simple and requires no programming knowledge.

The VIBE Interface

The screenshot displays the VIBE (Visual Bioinformatics Interface) software. The main workspace shows a workflow diagram with the following components:

- Input:** A 'Load' module (cylinder) feeds into a 'FASTA' module (cylinder).
- Search Engines:** The 'FASTA' module branches into three search engines: 'BLASTN', 'BLASTF', and 'Smith-Waterman' (all cylinders).
- Visualization:** Each search engine is connected to a corresponding 'SVView' module (cylinder) for visualization.

On the right side, there is a 'Details' panel titled 'Example - Search Comparison' with the following text:

Workspace Description:
 Here we show three different similarity search algorithms. BLAST is the fastest, but also least sensitive of the group. The FAST searches, a heuristic algorithm like BLAST, is slightly slower, but more sensitive. Smith-Waterman is a dynamic programming algorithm that guarantees the optimal (highest scoring) alignment between two sequences, but is computationally intensive.

IMPORTANT: Note that search modules in this pipeline may have used databases that are not available to you; replace these modules with new instances to allow VIBE to ask your server for its database.

At the bottom, a 'Execution Log' window shows the following entries:

```

May 25, 2006 1:01:55 AM COMPLETE: Type = FASTA Job Name = Search_1 Job ID = 1000000000
May 25, 2006 1:02:33 AM RUNNING: Type = SVView Job Name = View_1
May 25, 2006 1:02:37 AM COMPLETE: Type = SVView Job Name = View_1
May 25, 2006 1:02:39 AM RUNNING: Type = BLASTN Job Name = Search_2 Job ID = 1000000000
May 25, 2006 1:02:39 AM RUNNING: Type = SVView Job Name = View_2
    
```

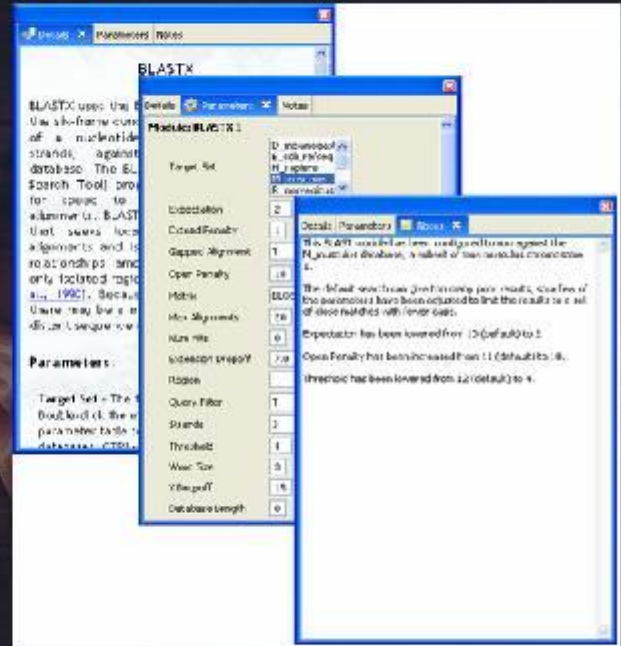
Slide notes

Common algorithms are represented as blocks or “modules” that can be dragged onto a workspace and connected to other algorithms already in place. Analysis results can be visualized to make sure that the pipeline is appropriate. Pipelines can then be saved, with or without data, and loaded later to run with new data or simply to review past results. Pipelines can also be emailed from within the application to collaborators who also use VIBE or stored in a central location where co-workers can access them.



Tool Information

- In-depth details on each tool and its parameters
- Easy-to-use table of parameters
- "Notes" page for user-annotation of an analysis



INCOGEN

Slide notes


Detailed information about each algorithm is available simply by clicking on the module that represents the algorithm. The parameters for each algorithm are easily accessible and easily changed. Specific notes can be added to any analysis module in a pipeline to give more information to others who may use the pipeline or to remind yourself later about which parameters you changed and why. A description can also be added to any workspace to add more information about the entire analysis pipeline.

Results Visualization

The screenshot displays two windows from the VIBE application. The left window shows a sequence alignment with a legend for 'PROGEM VIB 2.1' and 'VIB 2.1'. The right window shows a 'Restriction Enzyme Settings' dialog box with a table of enzymes and their recognition sites. The background of the slide features a person in a blue lab coat.

Select	Enzyme Name	Definition
<input type="checkbox"/>	BamHI	GGATC_C
<input type="checkbox"/>	BstI	AGATC_C
<input type="checkbox"/>	KpnI	ATPCS_AT
<input type="checkbox"/>	DraI	TTTAAA

- Interactive, graphical views of results
- Active links to external resources (e.g., Genbank)



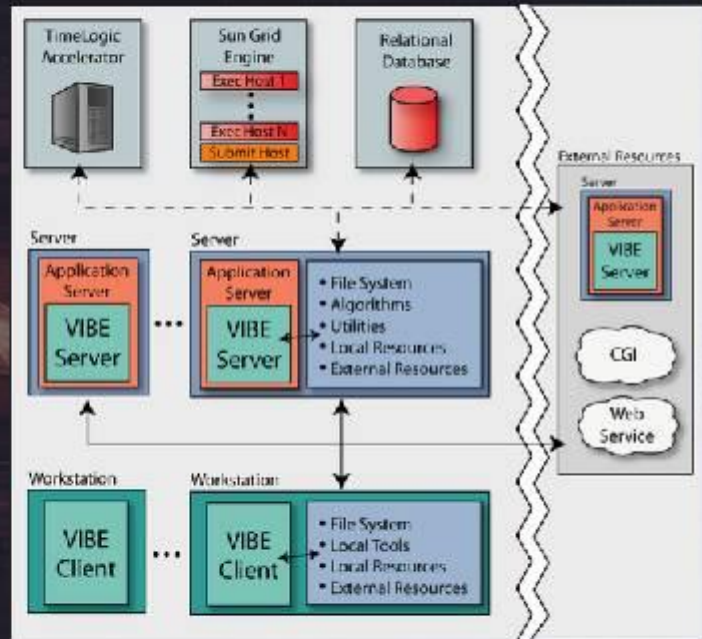
INCOGEN

Slide notes

Almost all data in VIBE can be visualized using one of the visualization modules provided with the VIBE application. The visualization modules provide an interactive interface so that you can remove or rearrange data or get more details about a particular piece of information. When appropriate, the viewers can link directly to external data sources, such as GenBank, to provide even more information about the visible data.

System Layout

- 100% Java
- OS-Independent
- Web-protocol (HTTP, SOAP, etc)
- Multi-server, multi-client, multi-resource
- Mix internal and external resources



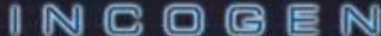

INCOGEN

Slide notes

VIBE is written in 100% Java, which makes it operating-system independent, and uses standard web-protocols. VIBE clients are capable of communicating with multiple VIBE servers, and multiple clients can communicate with a single server at the same time. VIBE servers broadcast their availability to clients, which allows users to easily choose a valid VIBE server from a list of all servers that are available. VIBE clients can also communicate with external resources, such as NCBI.

VIBE SDK

- Software Development Kit
- Allows integration of in-house or 3rd party tools into the VIBE client or server
 - Integration API (iAPI) and guides
 - External XML configuration



Slide notes

The VIBE platform can also be easily extended to incorporate other data types, algorithms, and viewers using the VIBE software development kit, or SDK.

VIBE as a Bioinformatics Teaching Platform

- Build on existing ease-of-use and extensive help features
- Allows students to explore many tools in biologically relevant contexts
- Concentrate on usage of tools and correct application
- Incorporate more sophisticated tracking capabilities, multi-media tutorial, etc.
- Feedback critical for success and subsequent stages



INCOGEN

Slide notes

Taken altogether, VIBE is an excellent bioinformatics teaching platform. It is easy to use and incorporates extensive help features. It encourages students to explore tools and resources in biologically relevant contexts so that they can concentrate on learning which tools should be used when, instead of using the one with which they are most familiar. VIBE also incorporates tracking capabilities and feedback mechanisms to facilitate future improvement. This feedback is critical to the future success of VIBE as a teaching platform.