The Computational Impact of Genomics on Biotechnology R&D (sort of…)

John "Scooter" Morris, Ph.D.

Genentech, Inc.

Biotechnology?

Means many things to many people

- Genomics
- Gene therapy
- Proteomics
- Diagnostics
- Drug delivery
- etc.

Biopharma – the use of biotechnology to produce pharmaceuticals

Genentech

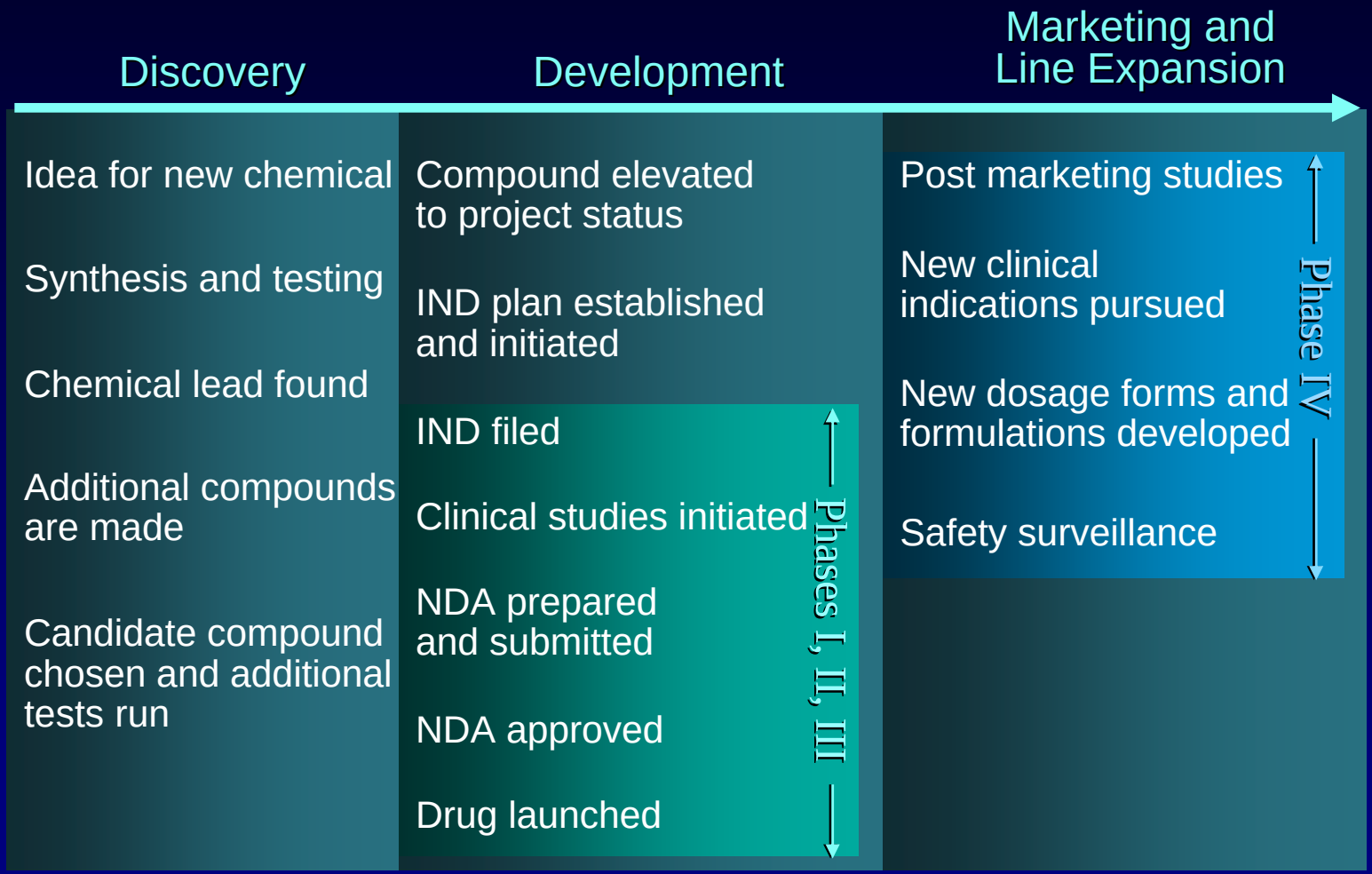
"Genentech is a pharmaceutical company dedicated to applying recombinant DNA technologies to unmet medical needs."

Founded 25 years ago

9 Marketed Products

- Human Growth Hormone Products
  - Protropin®, Nutropin®, NutropinAQ®, NutropinDepot™
- Activase®
- TNKase®
- Pulmozyme®
- Rituxan®
- Herceptin®

Clinical Development of Drugs



Discovery

From Craig Venter's slides:

Discovery won't wait

At Genentech, it will wait, but it will cost you…

$1 million / day

Discovery

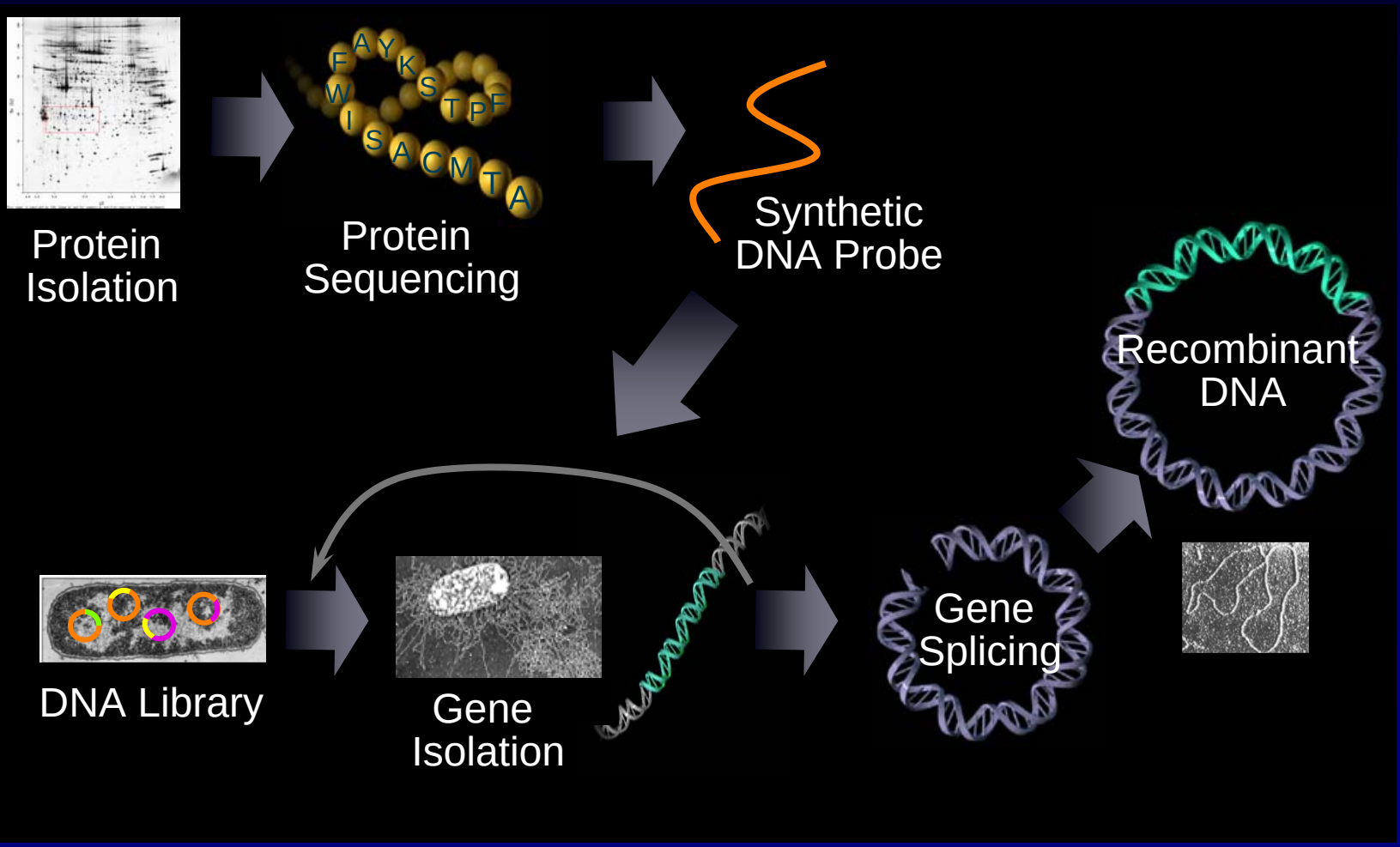
I'm going to focus on sequence analysis

Other aspects to Genentech's discovery program

- Basic research in diseases and disease states
- Animal models
- Clinical research
- "Humanized" Monoclonal Antibodies
- Protein structure determination
- Process sciences

All of these have their own computational needs

“Recombinant” Discovery (old)



“Recombinant” Discovery (old)

Process is very time consuming

- Months of experimentation and refining

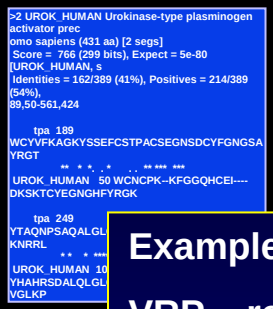
Process is error prone

- Assay development is expensive
- Assays may not specific enough
- Might get ambiguity from probe
- Might not get full length clone

Sequence database use

- Test against known proteins / DNA
- Help establish intellectual property

“Recombinant” Discovery (newer)



Order EST

PCR

Gene Splicing

Recombinant Plasmid DNA

Assay for activity

Example:

VRP – related to Genentech protein

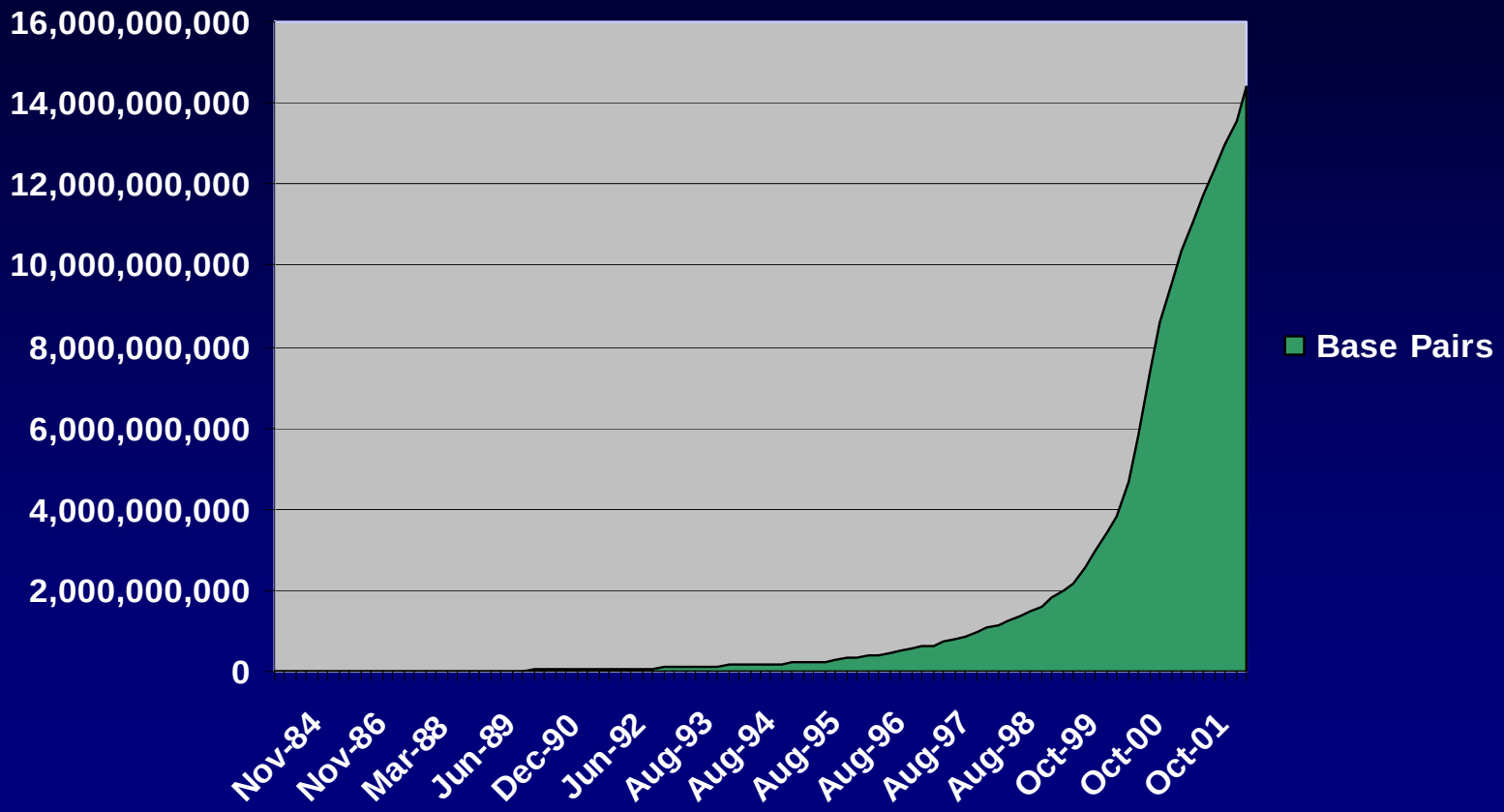
- **3 year research effort**
- **Found in early EST scan**

Comparison

Step	Get Protein	Get DNA	Full-Length Clone
OLD	Lab/Assay	Lab	Lab
	Months-years	Weeks-months	Weeks-months
NEW*	Select from database	Run program	Order for $25-$30
	Minutes	Minutes-hours	Minutes

* May still need to extend with PCR to get full length clone. Also still need to assay and express

Growth of Genbank



Similarity Searching

Proteins with similar function are similar

- Usually, this means the DNA is similar

Proteins with known function can be used as *probes* into database

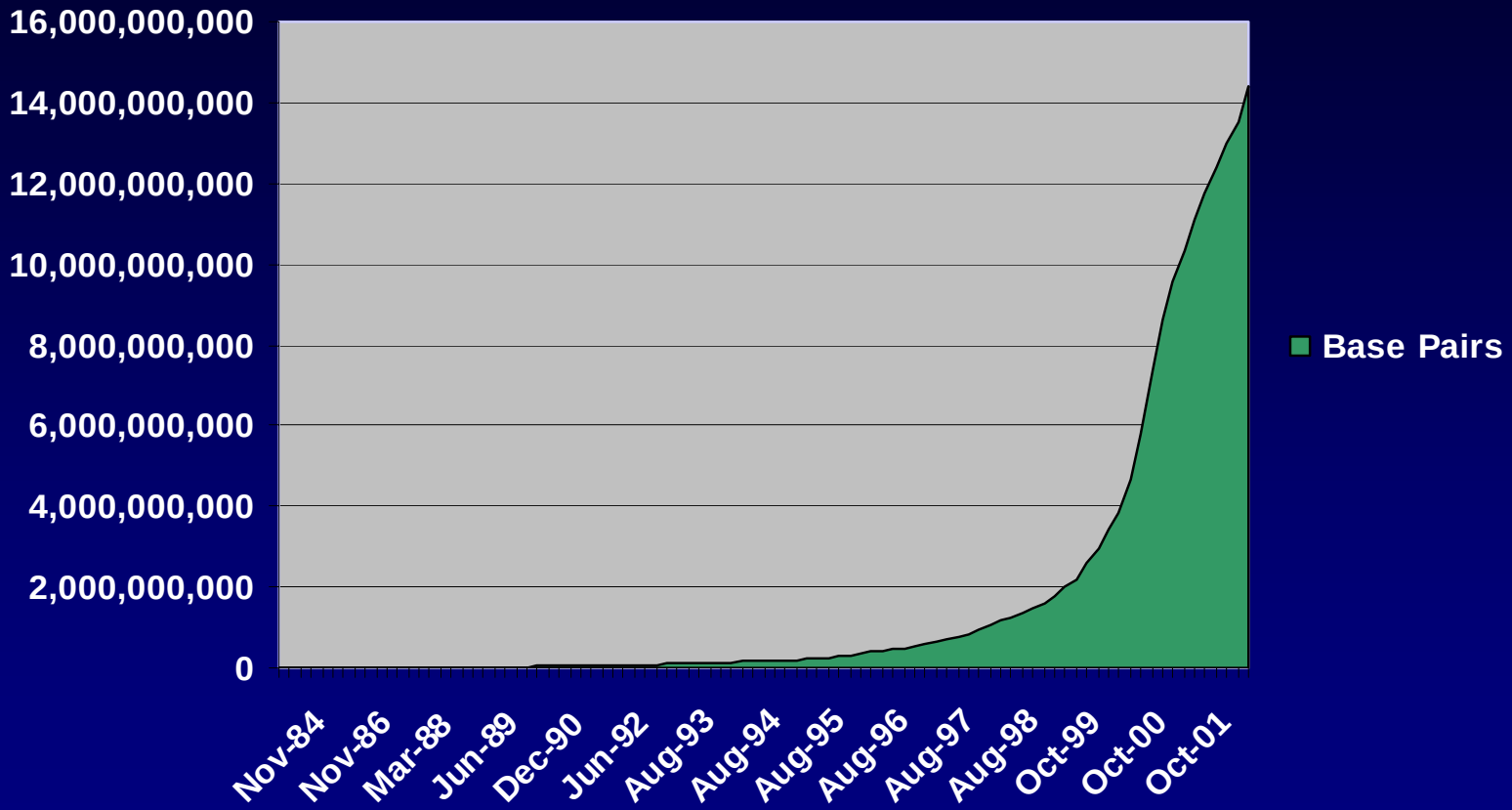
- Provides similar proteins, additional members of protein families
- Example: serine proteases

Main tool: *blast*

Blast

```
>2 UROK_HUMAN Urokinase-type plasminogen activator precursor /pid=CAA26268.1 –
   homo sapiens (431 aa) [2 segs]
 Score =  766 (299 bits), Expect = 5e-80 [UROK_HUMAN, seg 1/2]
 Identities = 162/389 (41%), Positives = 214/389 (54%), Gaps = 30/389 (7%), at 189,50-561,424

       tpa  189 WCYVFKAGKYSSEFCSTPACSEGNSDCYFGNGSAYRGTHSLTESGASCLPWNSMILIGKV
                **   *  *.  . *      . .  ** ***  ***  *    *  ******  .. .
 UROK_HUMAN   50 WCNCPK--KFGGQHCEI----DKSKTCYEGNGHFYRGKASTDTMGRPCLPWNSATVLQQT

       tpa  249 YTAQNPSAQALGLGKHNYCRNPDGDAKPWCHVLKNRRLTWEYCDVPSCS-----------
                * *    *  ************   .***.*     .   . * *  *.
 UROK_HUMAN  104 YHAHRSDALQLGLGKHNYCRNPDNRRRPWCYVQVGLKPLVQECMVHDCADGKKPSSPPEE

       tpa  298 ---TCGLRQYSQPQFRIKGGLFADIASHPWQAAIFAKHRRSPGERFLCGGILISSCWILS
                   ** ..  .*.*.* ** *  * . ** ***. .* *     ..*** *.* **..*
 UROK_HUMAN  164 LKFQCG-QKTLRPRFKIIGGEFTTIENQPWFAAIYRRH-RGGSVTYVCGGSLMSPCWVIS

       tpa  355 AAHCFQERFPPHHLTVILGRTYRVVPGEEEQKFEVEKYIVHKEFDDDT--YDNDIALLQL
                * ***.       * ***.     . * *****  *.**..  **   . ******..
 UROK_HUMAN  222 ATHCFIDYPKKEDYIVYLGRSRLNSNTQGEMKFEVENLILHKDYSADTLAHHNDIALLKI

       tpa  413 KSDSSRCAQESSVVRTVCLPPADLQLPDWTECELSGYGKHEALSPFYSERLKEAHVRLYP
                .*   **** *  ..*.***         * **..*.**   .    * *.**   *.*
 UROK_HUMAN  282 RSKEGRCAQPSRTIQTICLPSMYNDPQFGTSCEITGFGKENSTDYLYPEQLKMTVVKLIS

       tpa  473 SSRCTSQHLLNRTVTDNMLCAGDTRSGGPQANLHDACQGDSGGPLVCLNDGRMTLVGIIS
                   *   *     **  **** *     **    *.**********   ***** **.*
 UROK_HUMAN  342 HRECQQPHYYGSEVTTKMLCAAD-----PQWKT-DSCQGDSGGPLVCSLQGRMTLTGIVS

       tpa  533 WGLGCGQKDVPGVYTKVTNYLDWIRDNMR
                ** **  ** *****.*...* *** . .
 UROK_HUMAN  396 WGRGCALKDKPGVYTRVSHFLPWIRSHTK
```

Growth of Genbank



Computational Demands

Genbank has grown:

- 21,000X in 20 years
- 22X in the last 5 years

Significant growth in other public databases

- e.g. Swissprot, Procite, Blocks, Pfam

Advent of private databases

- e.g. Incyte, Celera

Other applications

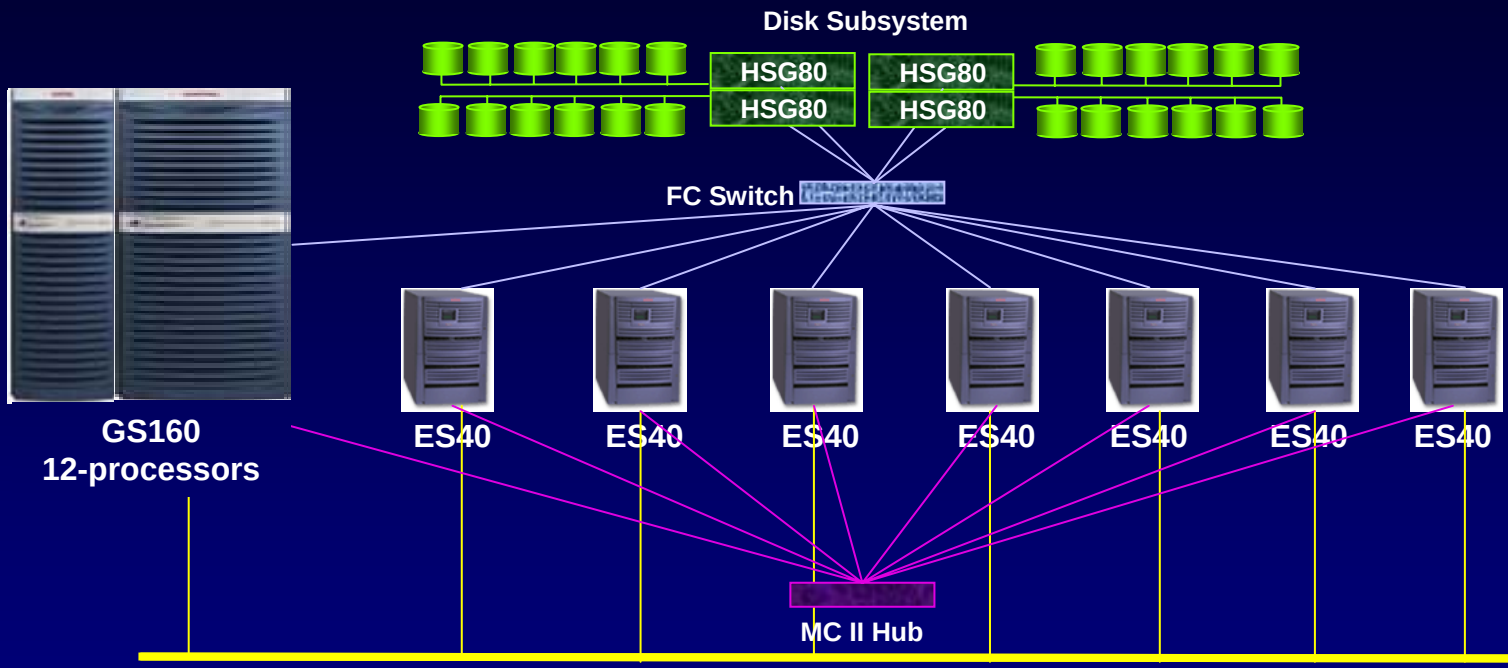
- Sequencing (both DNA and Protein)
- Microarray analysis
- High throughput screening
- Assay results

Computational Demands

Bioinformatics Job Mix

- Blast
  - I/O and integer intensive
  - Embarrassingly parallel
  - Large memory footprint
- Other applications (e.g. microarray analysis)
  - I/O & memory intensive
  - Floating point intensive
- User services
  - Web services
  - Appleshare
  - SAMBA
  - etc.

Bioinformatics Computing Evolution



Bioinformatics Computing

Approach was evolutionary

- Each step was an upgrade or an enhancement to existing computational resource
- Used existing tools whenever possible
- Maintain user expectations
- Minimize impact to discovery process

Current environment

- 1 GS160 (12-processors, 12GB)
- 7 ES40s (4-processors, 8GB)
- Can easily handle current normal blast demands
- Web interfaces to blast and other tools very popular
- Upgrading GS160 to handle additional microarray data
  - Protein-protein interaction studies
  - Floating point, CPU count intensive

Bioinformatics Computing

Why Alpha?

- Long history between Digital (now Compaq) and Genentech
- Wanted to take advantage of 64-bit address space
- Raw per-processor performance leader at the time
- Good I/O and floating point characteristics
- Excellent presence in biotechnology

Why Cluster?

- Substantially reduced database maintenance
  - One copy of the database
- Flexibility
  - Can migrate services as needed
- Ease of administration
  - Lots of users
  - Individual home directories
- Some increase in complexity
  - Getting services and filesystems right has taken some effort

Other Approaches

Large SMP systems
- 64 bit support
- Good I/O performance
- Generally poor price/performance
- Traditionally used at Genentech for computational chemistry and molecular modeling

Linux (IA32) clusters
- Excellent price/performance
- Particularly useful for back-end processing
- Must divide database up for large *blast* jobs
- Not as good for high I/O or floating point applications
- Pilot deployed at Genentech for *ab initio* calculations

Custom hardware
- Algorithm in firmware, PLAs, or ASICs
- Excellent performance
- Harder (impossible?) to adapt algorithms for local needs

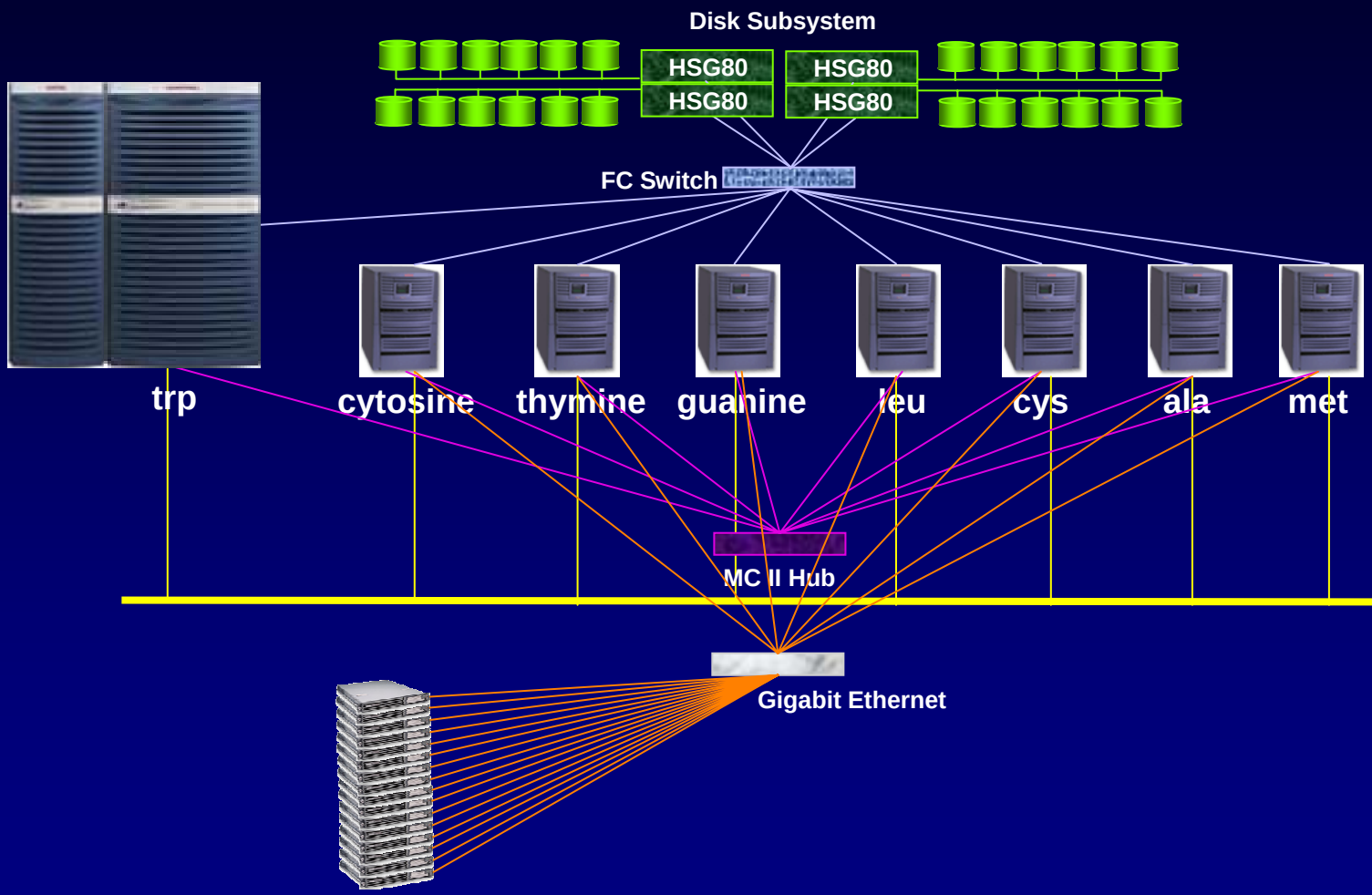
Futures Needs

Computational needs will continue to increase

- **Pharmacogenomics**
  - Personalized medicine
  - SNPs – Single nucleotide polymorphisms
- **Proteomics**
- **Searches for more distant homologs**
  - Human Genome: function of 42% of genes unknown
  - So, what *does* that 42% of genes code for?

How do we scale to meet future needs?

Bioinformatics Computing – Future?



Conclusions

Genentech's goal is to address unmet medical needs through recombinant DNA technology

- **Human therapeutics**

The availability of genomic data is dramatically reducing the time to discover medically relevant proteins

- **Quicker time to market**

It is also dramatically increasing our computational requirements …

- **… and increasing competitive pressures**

Conclusions

We've met our computing requirements (so far) through an evolutionary approach

Future computational needs will be much greater than today's

- **Proteomics**
- **Pharmacogenomics**
- **Functional genomics**

We hope to still be able to evolve to meet those needs

- **But we *will* meet the needs**

Acknowledgements

Colin Watanabe

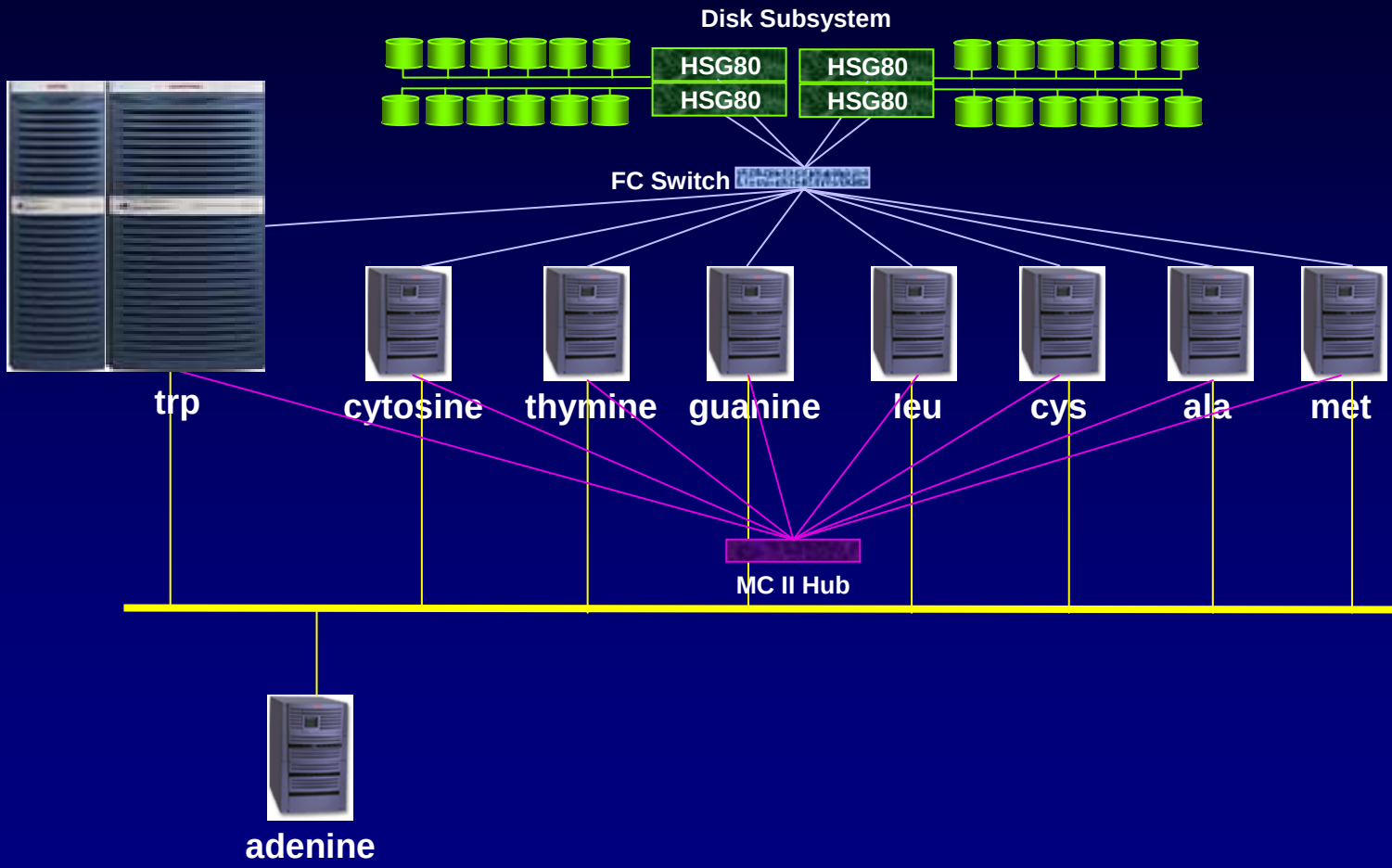
- Bioinformatics
- Molecular Biology

Carol Morita

- Molecular Biology

Questions?

Bioinformatics Computing



Bioinformatics Computing

Future directions

- Will look at Linux cluster after McKinley release
  - Still like 64 bit memory address
  - Clear price/performance leader for bioinformatics applications