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

- Idea for new chemical
- Synthesis and testing
- Chemical lead found
- Additional compounds are made
- Candidate compound chosen and additional tests run

Development

- Compound elevated to project status
- IND plan established and initiated
- IND filed
- Clinical studies initiated
- NDA prepared and submitted
- NDA approved
- Drug launched

Marketing and Line Expansion

- Post marketing studies
- New clinical indications pursued
- New dosage forms and formulations developed
- Safety surveillance

Phases I, II, III

Phase IV
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)

Protein Isolation → Protein Sequencing → Synthetic DNA Probe → Recombinant DNA

DNA Library → Gene Isolation → Gene Splicing
“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)**

Example:
- VRP – related to Genentech protein
  - 3 year research effort
  - Found in early EST scan

```plaintext
UROK_HUMAN 51 WCNCPK--KFGGQHCEI----DKSKTCYEGNGHFYRGK
YTAQIFQGAQALGLGKHNYCRNPDGDAKPWCHVLKRNRL
YHAHRSDALQLGLGKHNYCRNPDNRRRPWCYVQVGLKP

Example:
VRP – related to Genentech protein
```
# Comparison

<table>
<thead>
<tr>
<th>Step</th>
<th>Get Protein</th>
<th>Get DNA</th>
<th>Full-Length Clone</th>
</tr>
</thead>
<tbody>
<tr>
<td>OLD</td>
<td>Lab/Assay</td>
<td>Lab</td>
<td>Lab</td>
</tr>
<tr>
<td></td>
<td>Months-years</td>
<td>Weeks-months</td>
<td>Weeks-months</td>
</tr>
<tr>
<td>NEW*</td>
<td>Select from database</td>
<td>Run program</td>
<td>Order for $25-$30</td>
</tr>
<tr>
<td></td>
<td>Minutes</td>
<td>Minutes-hours</td>
<td>Minutes</td>
</tr>
</tbody>
</table>

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

- Base Pairs

- Timeline:
  - Nov-84 to Oct-01
  - Key dates:
    - Nov-84: 0 Base Pairs
    - Dec-90: 2,000,000,000 Base Pairs
    - Aug-97: 4,000,000,000 Base Pairs
    - Aug-99: 8,000,000,000 Base Pairs
    - Oct-00: 12,000,000,000 Base Pairs
    - Oct-01: 16,000,000,000 Base Pairs

- Scooter Morris, Genentech, Inc.
  (scooter@gene.com)
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

<table>
<thead>
<tr>
<th>tpa</th>
<th>WCYVFKAGKYSSEFCSTPACSEGNSDCYFGNOSAYRTSGLTSEGASCLFWNSMLIGKV</th>
</tr>
</thead>
<tbody>
<tr>
<td>UROK_HUMAN</td>
<td>50 WCNCPK--KFGQHCEI-----DEKSTCYEGNHFYGKASTDMGRPLFWNSATVLQQT</td>
</tr>
<tr>
<td>tpa</td>
<td>YTAQNPASQALGLKGNCRNPDGDAGFCHVLKNRRNRTWYCDVFSCS------------</td>
</tr>
<tr>
<td>UROK_HUMAN</td>
<td>104 YAHARSDALQGLKHCNRCNPNNRRPWCYQVGLKPLQECMVHDCADGKPSPPPE</td>
</tr>
<tr>
<td>tpa</td>
<td>---TCGLRQYSPQFQRIKGLFAHSLHPWAQAAIFAKHRRSPGERFLCGGILISWILS</td>
</tr>
<tr>
<td>UROK_HUMAN</td>
<td>164 LKFQCG-QKTLRPRFHIIGEEFTTIENQPWFAAIYRRH-RGGSVTVCGGSLMSFCWIS</td>
</tr>
<tr>
<td>tpa</td>
<td>AAIHCDFQERFPPHHTLTINGRTVRYPQEEQKFEVEKHY1HEPDD---YDNDIALQL</td>
</tr>
<tr>
<td>UROK_HUMAN</td>
<td>222 ATHCFIDYPKEDTLYGLSRLNSNTQGMKFEVENLILHDYSADTLLAHNDIALKI</td>
</tr>
<tr>
<td>tpa</td>
<td>KSDSSRCQESSSLVTVQPLFADLQLPDWTECGLSUGKHEALSPFLERLKEAHVRTYYP</td>
</tr>
<tr>
<td>UROK_HUMAN</td>
<td>282 RSIKEGRCQPSRTIQICLPSMNYDPQFSGTCEIGFTGKENSTDLYPEQLKVTVKLIS</td>
</tr>
<tr>
<td>tpa</td>
<td>SSCRCTSQHLLNRTVTDNLMLRAGDTRSGPQANLDACQGDSGGLVCLNDGRMLTVGIIS</td>
</tr>
<tr>
<td>UROK_HUMAN</td>
<td>342 HRECQQPHYGSEVTMCLAAA----DQWKT-DSCQGDSGSLVCLQGRTLTGIVS</td>
</tr>
<tr>
<td>tpa</td>
<td>WGLGCGQKDVPGVYTKTYNLDWIRDNMR</td>
</tr>
<tr>
<td>UROK_HUMAN</td>
<td>396 WGRGCALDKPQVTRVSHFLPWIRSHTK</td>
</tr>
</tbody>
</table>
Growth of Genbank

Base Pairs


0 2,000,000,000 4,000,000,000 6,000,000,000 8,000,000,000 10,000,000,000 12,000,000,000 14,000,000,000 16,000,000,000

Scooter Morris, Genentech, Inc.
scooter@gene.com
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

GS160
12-processors

Disk Subsystem

HSG80
HSG80
HSG80
HSG80

FC Switch

MC II Hub

ES40
ES40
ES40
ES40
ES40
ES40
ES40
ES40

SC2001
November 13, 2001  page 17
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?

Disk Subsystem
HSG80
HSG80
HSG80
HSG80

FC Switch

MC II Hub

Gigabit Ethernet

cytosine
thymine
leu
cys
ala
met
trp

Scooter Morris, Genentech, Inc.
(scooter@gene.com)
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

Future directions

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