Introduction of opportunity and challenge in Biostatistics and Bioinformatics to Math major students

George C. Tseng

Department of Biostatistics

Department of Human Genetics

University of Pittsburgh

Outline

- Possible applications of probability and statistics
- Biostatistics
  - Academic research
  - Industry
- Bioinformatics
- Transitions
  - Application => Ph.D. student => Research/Job
- Some final words
  - Curriculum and preparation
  - Studying abroad??

My CV

93-97	B.S. Mathematics	National Taiwan Univ.
97-99	M.S. Statistics	National Taiwan Univ.
99-00	Statistics	UCLA
00-03	Ph.D. Biostatistics	Harvard University
03~	Biostatistics	University of Pittsburgh

Brain		Income
High	Mathematician	Low
	Applied Mathematician	
	Statistician	
	Biostatistician	
	Epidemiologist	
	Physician	
Low		High

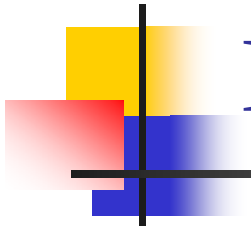
I. Applications of statistics

- Agricultural science
- Social science: education, psychology,…
- Financial mathematics
- Actuarial science
- Biomedical science
  - Biostatistics, medical imaging, Biomath, Biophysics...
  - Bioinformatics, Computational Biology

……

II. Biostatistics

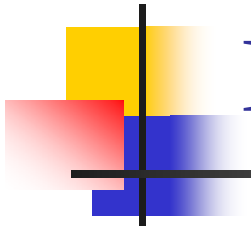
- Statistical research usually motivated by applications of public health, medicine or genetics.
- Research results should at least have one area of application.
- Harvard, Johns Hopkins, U Washington, U North Carolina-Chapel Hill, U Michigan, U Minnesota, U Pittsburgh, Case Western Reserve Univ., Columbia Univ., Emory Univ., Boston Univ., UCLA, U Wisconsin-Madison



II. Biostatistics

Research Areas: (from the dept. website)

- **Dept. of Biostatistics at Harvard**
  - **AIDS research**
  - **Cancer research**
  - **Computational biology & Bioinformatics**
  - **Environmental statistics**
  - **Genetic epidemiology**
  - **Neurostatistics**
  - **Psychiatric biostatistics**



II. Biostatistics

Research Areas: (from the dept. website)

- **Dept. of Biostatistics at Univ. Pittsburgh**
  - **Cancer treatment trials**
  - **Health outcomes/health services research**
  - **Environmental & occupational epidemiology**
  - **Radiological imaging system**
  - **Psychiatric research**
  - **Computational biology & Bioinformatics**
  - **Statistical methodology**

II. Biostatistics

A simple example of survival analysis:

A new drug and an old drug are applied to cancer patients. Survival time of each patients are recorded after treatment. The study was terminated at 60 months.

ID	group	relapse	survival
1	1	1	12
2	1	0	60
3	1	0	60
4	1	0	60
5	1	1	12
6	1	0	60
7	1	0	60
8	1	0	60
9	1	0	60
10	1	0	60
11	0	1	1
12	1	0	60

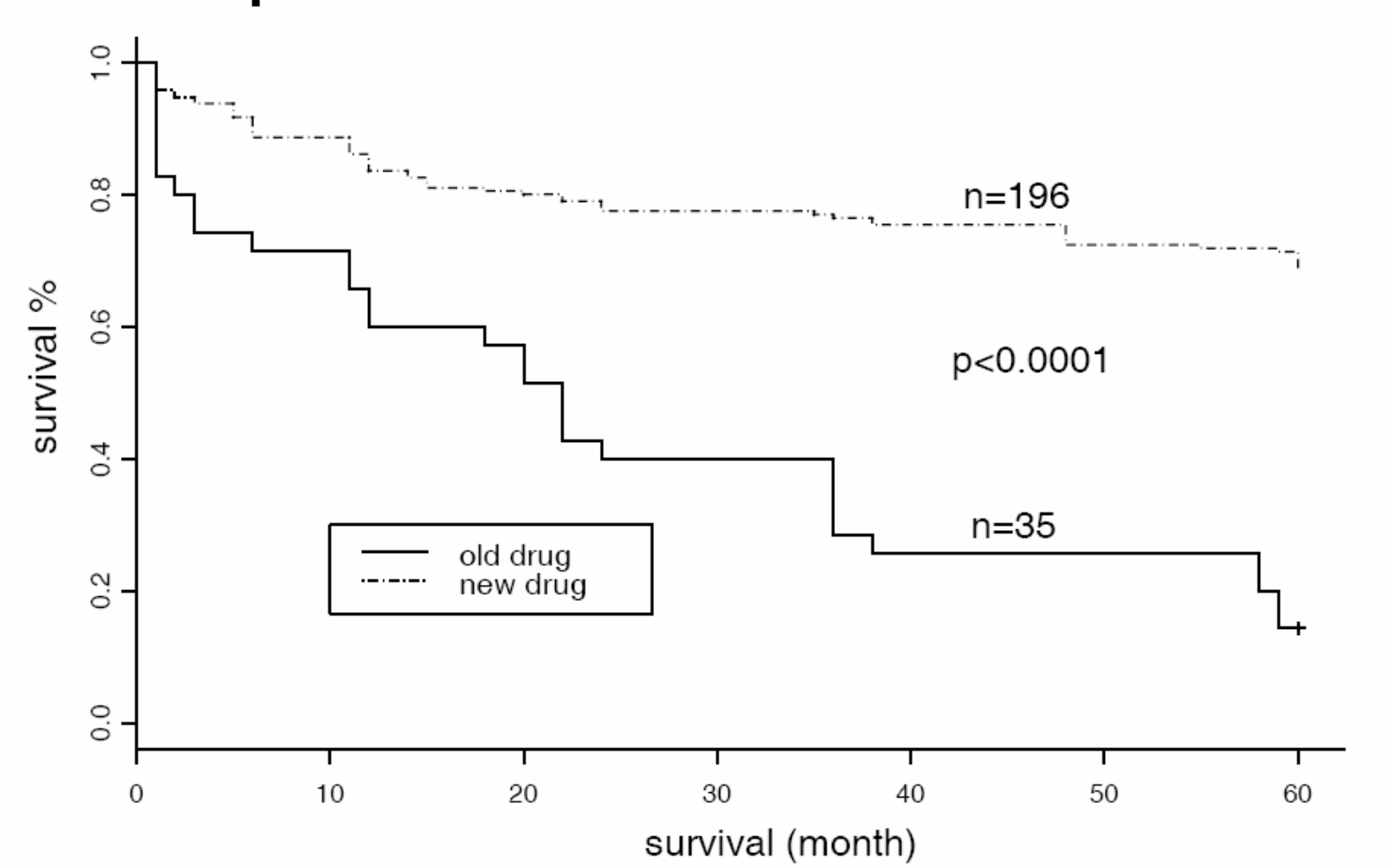
New drug (1):196 patients
Old drug (0): 35 patients
Relapse (1): died
Relapse (0): survived over

Q: How do we rigorously and confidently decide the new drug is better than the old drug?

II. Biostatistics

A simple example of survival analysis:

Kaplan-Meier curve



II. Biostatistics

A simple example of survival analysis:

- Compare the difference of two survival curves.
- Modelling censoring and survival model.
  - Early drop out patients
  - Patients participate in interim of study
- Experimental design
  - Case-control matched study
  - Early termination

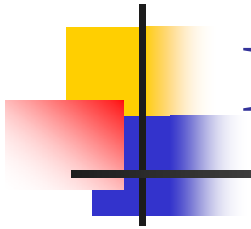
II. Biostatistics

Employment of alumni (Dept. of Biostatistics, Univ. of Pittsburgh)

Type of Employment	M.S./M.P.H.	Ph.D/Sc.D.
Academic Institutions	52	48
Government Agencies	28	11
Other Health Research Groups*	25	4
Private Industry	31	23
Other (includes students continuing for doctoral degree)	20	2
Unknown	23	7
Deceased	2	1
Total	181	96

II. Biostatistics: working in university

- **Tenure track**
  - **Research (publication and academic activity)**
    - **Methodology research**
    - **Collaborative research**
  - **Teaching**
  - **Grant proposals**
  - **Service (committees, advising students…)**
- **Research track**
  - **Research**
    - **Collaborative**
    - **Methodology**
  - **Grant proposals**



II. Biostatistics: working in government

Centers for Disease Control
National Institutes of Health
U.S. Census Bureau
National Center for Health Statistics
Food and Drug Administration

II. Biostatistics:
working in pharmaceutical company

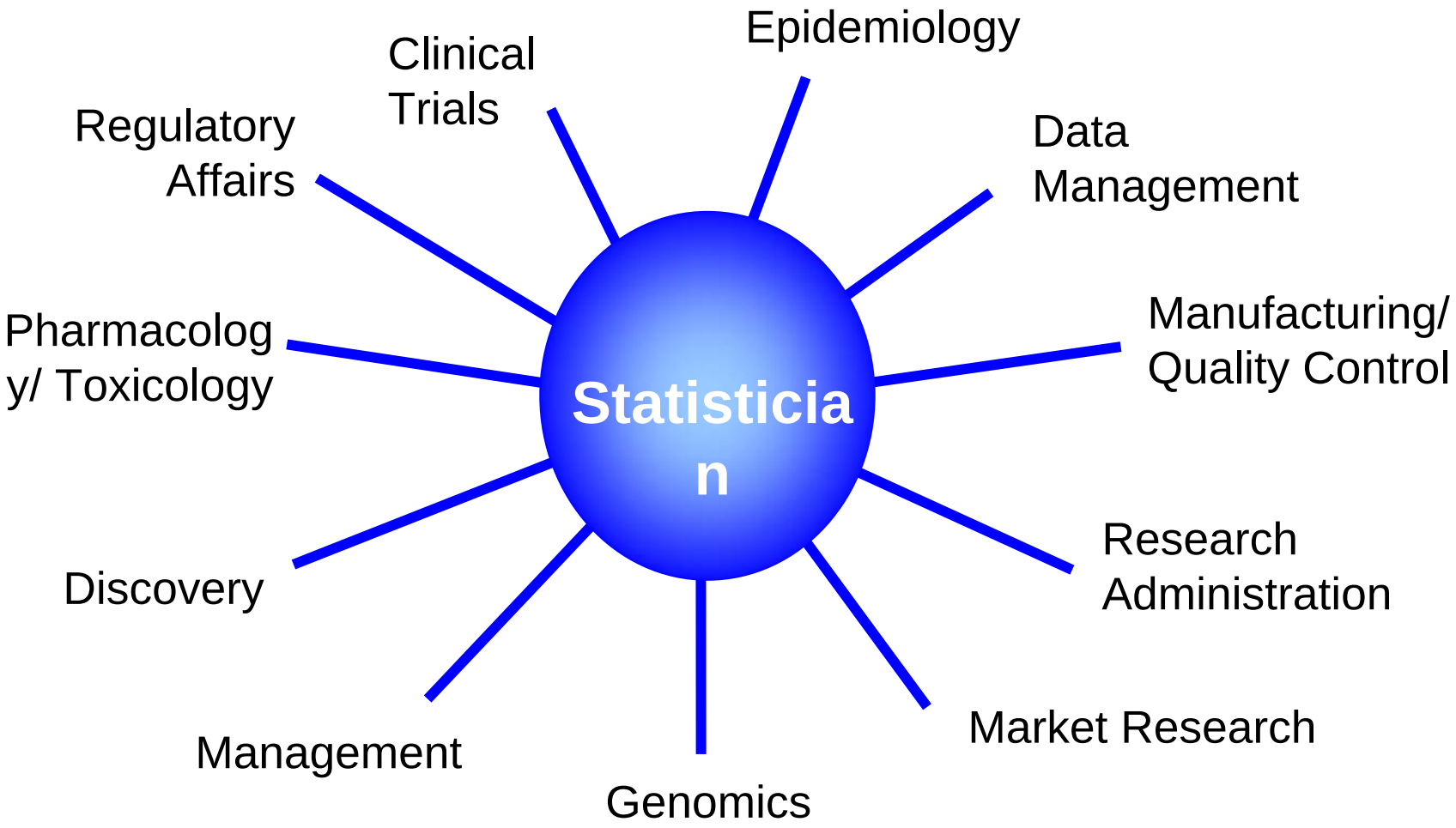
Merck: one of the largest drug companies in the US

- Global, research-driven pharmaceutical company
  - ~62,000 employees worldwide in 26 countries
  - In 2004, $22.9 billion in sales, $5.8 billion in net income, $3 billion invested in research
- Broad range of products
- Ranked in “100 Best to Work For” and “America’s Most Admired” and “Global Most Admired”

Info. from Merck & Co., Inc.

II. Biostatistics:
working in pharmaceutical company

Areas of Application



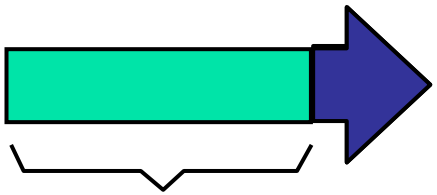
Info. from Merck & Co., Inc.

II. Biostatistics:
working in pharmaceutical company

New Drug Development

Creation	Role of Statistician
▪ Drug discovery ▪ Chemical synthesis ▪ Laboratory testing ▪ Animal testing ▪ Formulation of ingredients	▪ Analyze high throughput screening results ▪ Design screening strategies and select analogs ▪ Analyze dose-response studies ▪ Employ bioassay techniques ▪ Evaluate carcinogenic potential ▪ Evaluate reproductive and genetic toxicology

(2 - 4 Years)

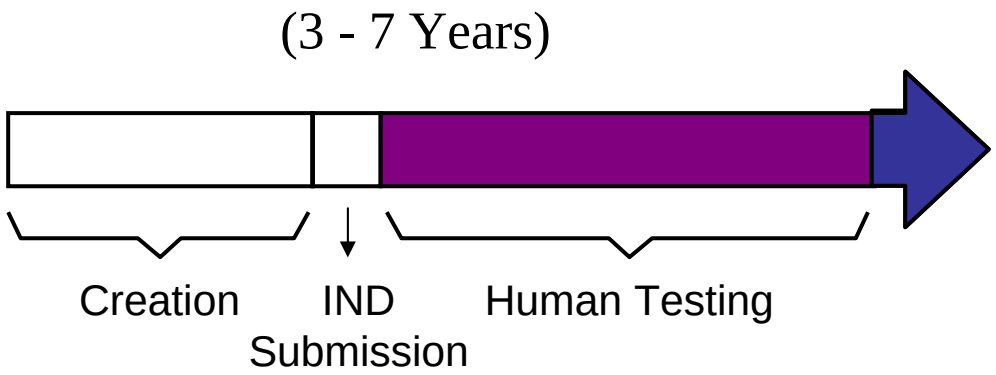


Info. from Merck & Co., Inc.

II. Biostatistics:
working in pharmaceutical company

New Drug Development

Human Testing	Role of Statistician
■ Phase I - Safety	■ Propose statistical methodology
■ Phase II a - Proof of Concept	■ Approve study protocols
■ Phase II b - Dose-Ranging	■ Interact with Project Team
■ Phase III - Safety and Efficacy	■ Analyze and interpret early studies



Info. from Merck & Co., Inc.

II. Biostatistics:
working in pharmaceutical company

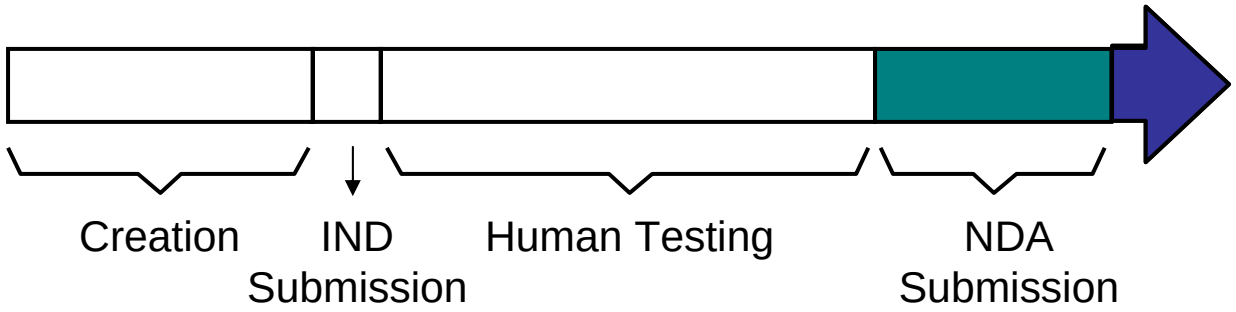
New Drug Development

New Drug Application	Role of Statistician

- FDA submission and review
- New drug available to patients and physicians

(1 - 3 years to prepare,
1 year to review)

- Summarize across studies
- Prepare statistical technical section
- Present methodology and results to FDA

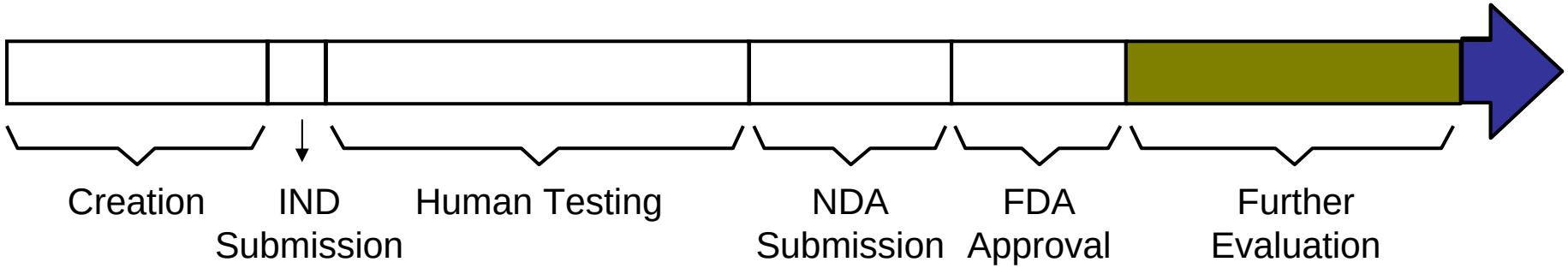


Info. from Merck & Co., Inc.

II. Biostatistics:
working in pharmaceutical company

New Drug Development

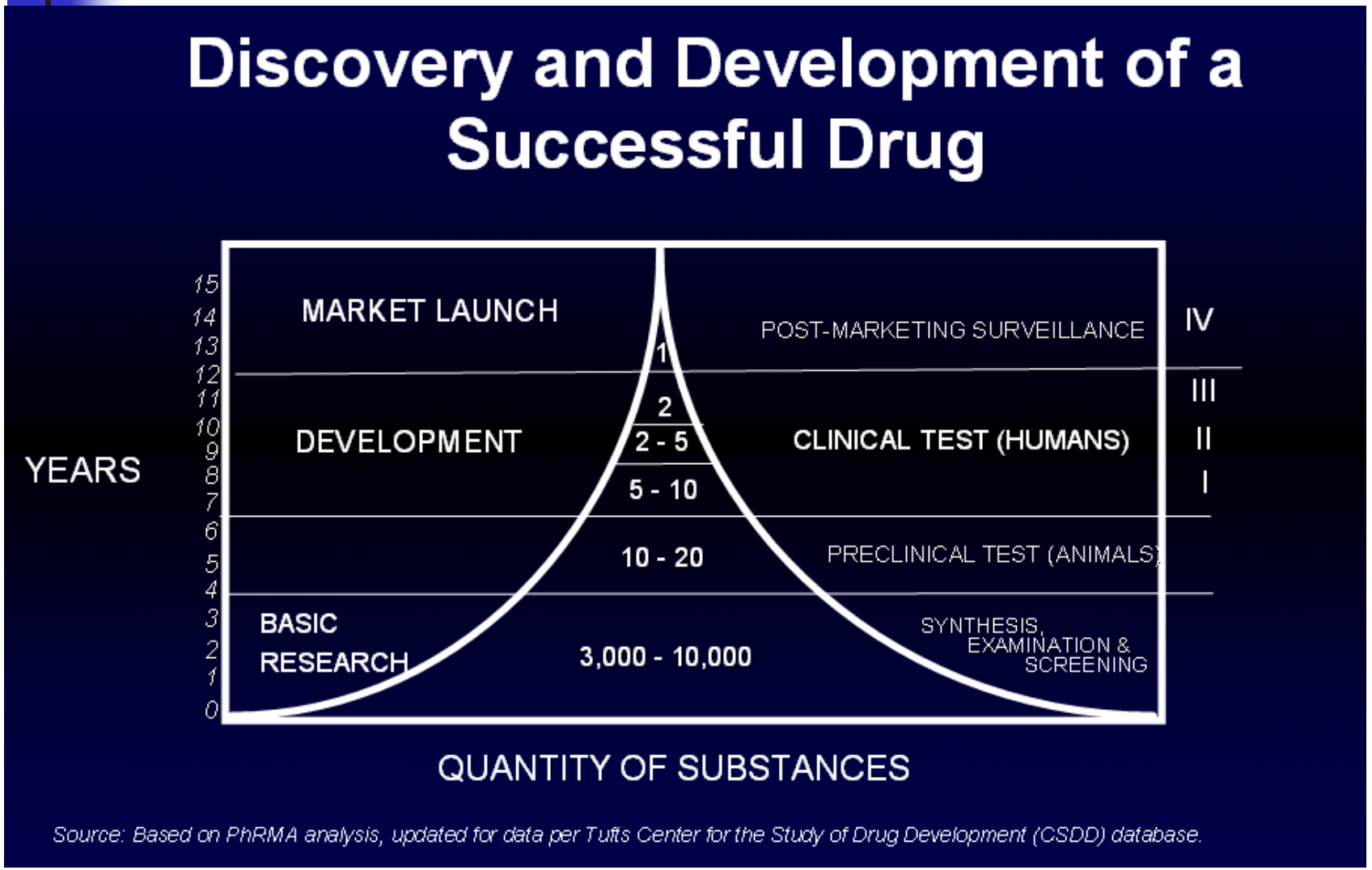
Further Evaluation	Role of Statistician
▪ Ongoing ▪ Additional uses ▪ Additional side effects ▪ Modification of dosage or form	▪ Design and analyze post-marketing studies ▪ Submit papers for publication



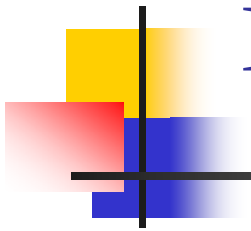
Info. from Merck & Co., Inc.

II. Biostatistics:

working in pharmaceutical company



Info. from Merck & Co., Inc.



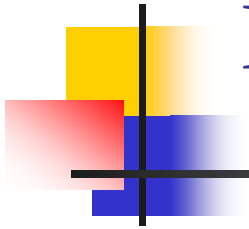
III. Bioinformatics

A simple example of motif finding

Combinatorial Gene Regulation

- A microarray experiment showed that when gene X is knocked out, 20 other genes are not expressed
  - **How can one gene have such drastic effects?**

From http://www.bioalgorithms.info/



III. Bioinformatics

A simple example of motif finding

Regulatory Proteins

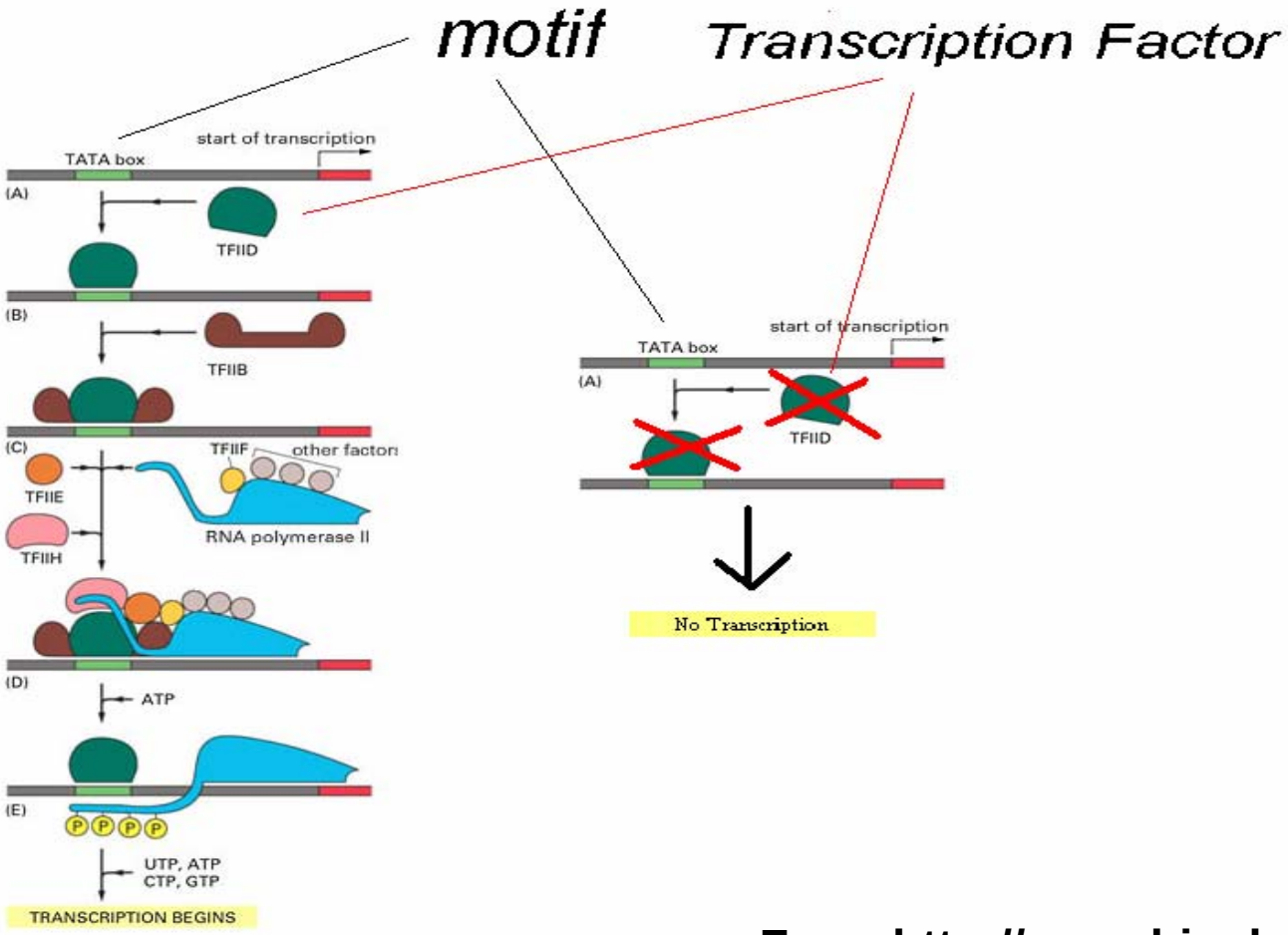
- Gene X encodes regulatory protein, a.k.a. a ***transcription factor*** (TF)
- The 20 unexpressed genes rely on gene X's TF to induce transcription
- A single TF may regulate multiple genes

From http://www.bioalgorithms.info/

III. Bioinformatics

A simple example of motif finding

Transcription Factors and Motifs



From http://www.bioalgorithms.info/

III. Bioinformatics

A simple example of motif finding

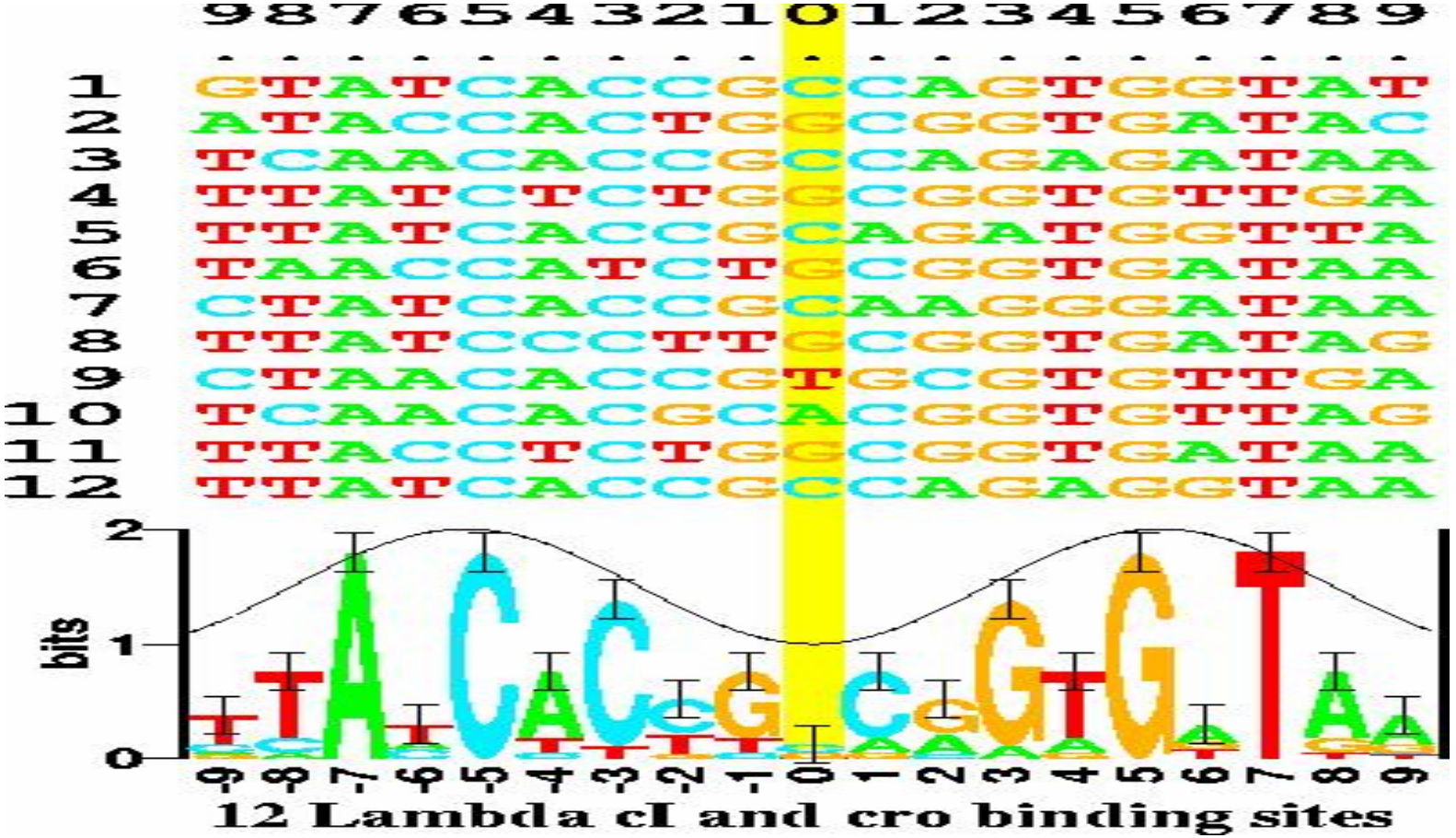
Motifs and Transcriptional Start Sites

From http://www.bioalgorithms.info/

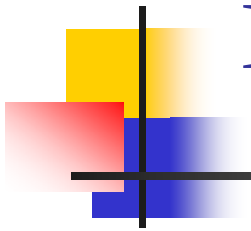
III. Bioinformatics

A simple example of motif finding

Motif Logos: An Example



From http://www.bioalgorithms.info/



III. Bioinformatics

A simple example of motif finding

Random Sample

atgaccgggatactgataccgtatttggcctaggcgtacacattagataaacgtatgaagtacgttagactcggcgccgccg

acccctatttttgagcagatttagtgacctggaaaaaaaatttgagtacaaaacttttccgaatactgggcataaggtaca

tgagtatccctgggatgacttttgggaacactatagtgctctcccgattttgaatatgtaggatcattcgccagggtccga

gctgagaattggatgaccttgtaagtgttttccacgcaatcgcgaaccaacgcggacccaaaggcaagaccgataaaggaga

tccctttgcggtaatgtgccgggaggctggttacgtagggaagccctaacggacttaatggcccacttagtccacttatag

gtcaatcatgttcttgtgaatggattttaactgagggcatagaccgcttggcgcacccaaattcagtgtgggcgagcgcaa

cggttttggcccttgttagaggcccccgtactgatggaaactttcaattatgagagagctaatctatcgcgtgcgtgttcat

aacttgagttggtttcgaaaatgctctggggcacatacaagaggagtcttccttatcagttaatgctgtatgacactatgta

ttggcccattggctaaaagcccaacttgacaaatggaagatagaatccttgcatttcaacgtatgccgaaccgaaagggaag

ctggtgagcaacgacagattcttacgtgcattagctcgcttccggggatctaatagcacgaagcttctgggtactgatagca

III. Bioinformatics

A simple example of motif finding

Implanting Motif **AAAAAAAGGGGGGG**

atgaccgggatactgatAAAAAAAAGGGGGGGggcgtacacattagataaacgtatgaagtacgttagactcggcgccgccg

acccctatttttgagcagatttagtgacctggaaaaaaatttgagtacaaaacttttccgaataAAAAAAAAGGGGGGGa

tgagtatccctgggatgacttAAAAAAAAGGGGGGGtgctctcccgattttgaatatgtaggatcattcgccagggtccga

gctgagaattggatgAAAAAAAAGGGGGGGtccacgcaatcgcgaaccaacgcggacccaaaggcaagaccgataaaggaga

tccctttgcggtaatgtgccgggaggctggttacgtagggaagccctaacggacttaatAAAAAAAAGGGGGGGcttatag

gtcaatcatgttcttgtgaatggatttAAAAAAAAGGGGGGGgaccgcttggcgcacccaaattcagtgtgggcgagcgcaa

cggttttggcccttgttagaggcccccgtAAAAAAAAGGGGGGGcaattatgagagagctaatctatcgcgtgcgtgttcat

aacttgagttAAAAAAAAGGGGGGGctggggcacatacaagaggagtcttccttatcagttaatgctgtatgacactatgta

ttggcccattggctaaaagcccaacttgacaaatggaagatagaatccttgcatAAAAAAAAGGGGGGGaccgaaagggaag

ctggtgagcaacgacagattcttacgtgcattagctcgcttccggggatctaatagcacgaagcttAAAAAAAAGGGGGGGa

III. Bioinformatics

A simple example of motif finding

Where is the Implanted Motif?

atgaccgggatactgatAAAAAAAAGGGGGGGgcgtacacattagataaacgtatgaagtacgttagactcggcgccgccg

acccctatttttgagcagatttagtgacctggaaaaaaaatttgagtacaaaacttttccgaataAAAAAAAAGGGGGGGa

tgagtatccctgggatgacttAAAAAAAAGGGGGGGtgctctcccgatttttgaatatgtaggatcattcgccagggtccga

gctgagaattggatgAAAAAAAAGGGGGGGtccacgcaatcgcgaaccaacgcggacccaaaggcaagaccgataaaggaga

tccctttgcggtaatgtgccgggaggctggttacgtagggaagccctaacggacttaatAAAAAAAAGGGGGGGcttatag

gtcaatcatgttcttgtgaatggatttAAAAAAAAGGGGGGGgaccgcttggcgcacccaaattcagtgtgggcgagcgcaa

cggttttggcccttgttagaggcccccgtAAAAAAAAGGGGGGGcaattatgagagagctaatctatcgcgtgcgtgttcat

aacttgagttAAAAAAAAGGGGGGGctggggcacatacaagaggagtcttccttatcagttaatgctgtatgacactatgta

ttggcccattggctaaaagcccaacttgacaaatggaagatagaatccttgcatAAAAAAAAGGGGGGGaccgaaagggaag

ctggtgagcaacgacagattcttacgtgcattagctcgcttccggggatctaatagcacgaagcttAAAAAAAAGGGGGGGa

III. Bioinformatics

A simple example of motif finding

Implanting Motif **AAAAAAGGGGGGG** with Four Mutations

atgaccgggatactgatAgAAgAAAGGttGGGggcgtacacattagataaacgtatgaagtacgttagactcggcgccgccg

acccctatttttgagcagatttagtgacctggaaaaaaatttgagtacaaaacttttccgaatacAAtAAAAcGGcGGGa

tgagtatccctgggatgacttAAAAtAAtGGaGtGGtgctctcccgattttgaatatgtaggatcattcgccagggtccga

gctgagaattggatgcAAAAAAAGGGattGtccacgcaatcgcgaaccaacgcggacccaaaggcaagaccgataaaggaga

tccctttgcggtaatgtgccgggaggctggttacgtagggaagccctaacggacttaatAtAAtAAAGGaaGGGcttatag

gtcaatcatgttcttgtgaatggatttAAcAAtAAGGGctGGgaccgcttggcgcacccaaattcagtgtgggcgagcgcaa

cggttttggcccttgttagaggcccccgtAtAAAcAAGGaGGGccaattatgagagagctaatctatcgcgtgcgtgttcat

aacttgagttAAAAAAtAGGGaGccctggggcacatacaagaggagtcttccttatcagttaatgctgtatgacactatgta

ttggcccattggctaaaagcccaacttgacaaatggaagatagaatccttgcatActAAAAAGGaGcGGaccgaaagggaag

ctggtgagcaacgacagattcttacgtgcattagctcgcttccggggatctaatagcacgaagcttActAAAAAGGaGcGGa

III. Bioinformatics

A simple example of motif finding

Where is the Motif???

atgaccgggatactgatagaagaaaggttgggggcgtacacattagataaacgtatgaagtacgttagactcggcgccgccg

acccctatttttgagcagatttagtgacctggaaaaaaaatttgagtacaaaacttttccgaatacaataaaacggcggga

tgagtatccctgggatgacttaaaataatggagtggtgctctcccgatttttgaatatgtaggatcattcgccagggtccga

gctgagaattggatgcaaaaaaagggattgtccacgcaatcgcgaaccaacgcggacccaaaggcaagaccgataaaggaga

tccctttgcggtaatgtgccgggaggctggttacgtagggaagccctaacggacttaatataataaaggaagggcttatag

gtcaatcatgttcttgtgaatggatttaacaataagggctgggaccgcttggcgcacccaaattcagtgtgggcgagcgcaa

cggttttggcccttgttagaggcccccgtataaacaaggagggccaattatgagagagctaatctatcgcgtgcgtgttcat

aacttgagttaaaaaatagggagccctggggcacatacaagaggagtcttccttatcagttaatgctgtatgacactatgta

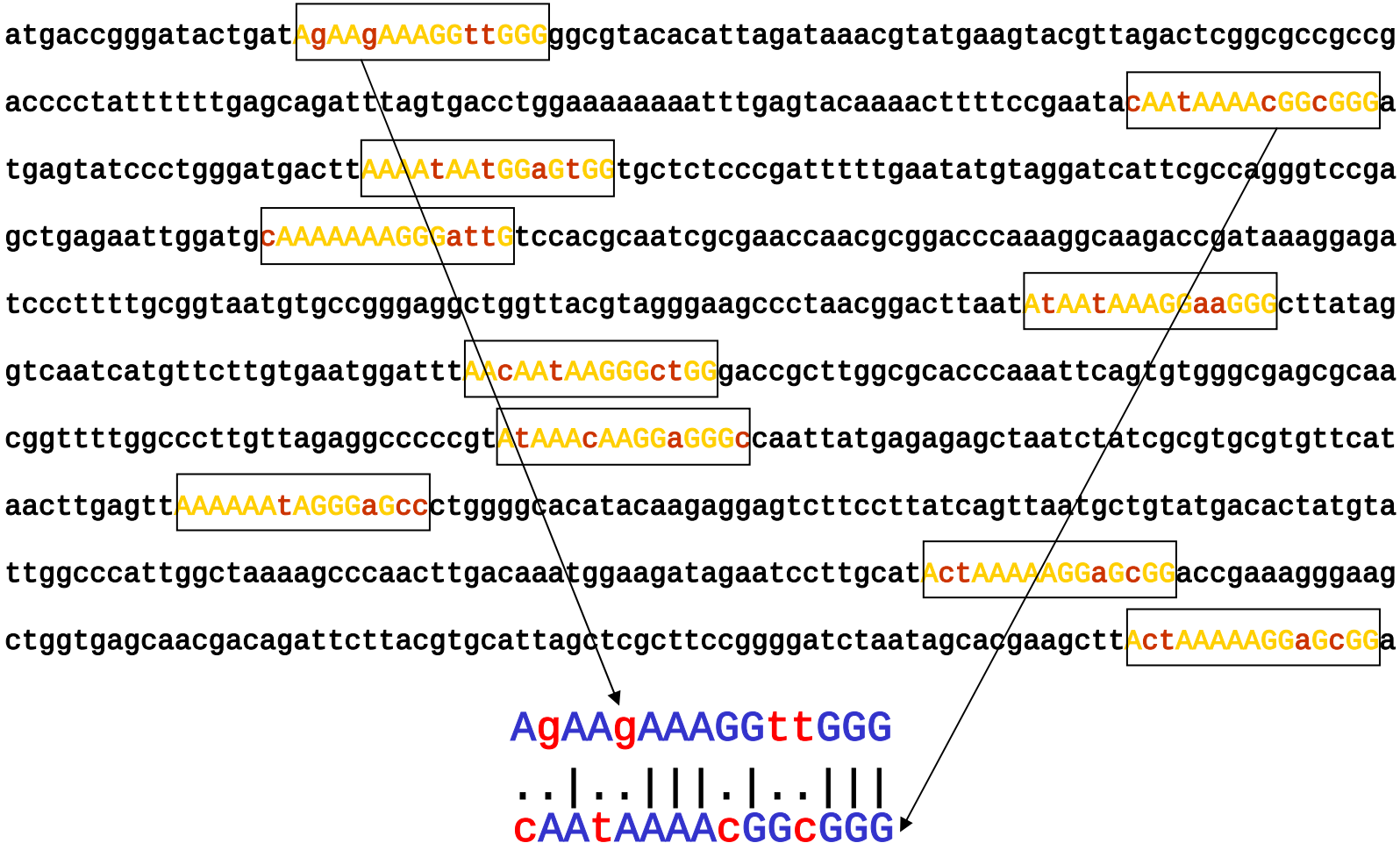
ttggcccattggctaaaagcccaacttgacaaatggaagatagaatccttgcatactaaaaaggagcggaccgaaagggaag

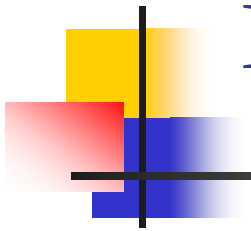
ctggtgagcaacgacagattcttacgtgcattagctcgcttccggggatctaatagcacgaagcttactaaaaaggagcgga

III. Bioinformatics

A simple example of motif finding

Why Finding (15,4) Motif is Difficult?





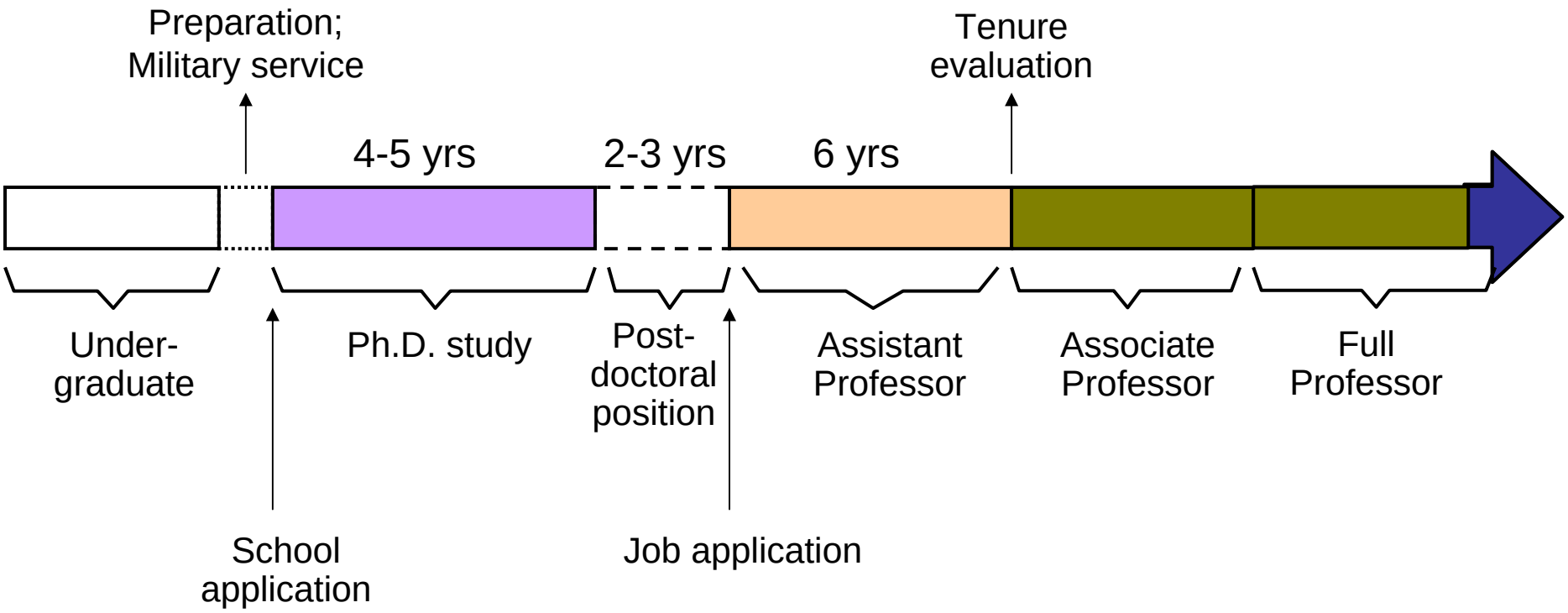
III. Bioinformatics

A simple example of motif finding

Questions:

- How to develop a good probabilistic model for the motifs?
- Is the computation affordable to search the whole genome? (Human genome is around 3 billion base pair long.)
- How to evaluate the statistical significance of the motifs you find?

IV. Transitions



IV. Transitions: Application

- GRE, TOEFL, GPA
- **Recommendation letter**
- **Study plan**

- Prepare and ask around early: take GRE and TOEFL; identify professors for recommendation letters and advises
- Academia Sinica (a good place to stay for short term transition and preparation)

IV. Transitions: Ph.D. study

- Settle down and enjoy
- Improve English; think open and American
- Professor, classmates, office-mates, colleagues are good assets for your future

Financial situation:

- Stipend (US$1600-$300tax) from TA or RA
- Rent US$400~500. Living cost $300~500.

IV. Transitions: Research/Job

- Going to academic is usually more busy than going to industry but with more freedom.
- No boss v.s. with a boss
- Irregular/flexible working hour v.s. regular working hour

?? $$$??

IV. Transitions: Research/Job

University (9 months)

Institution Type	Title	Years in Rank	Count	Median	3rd Quartile	90th Percentile
Research University	Assistant Professor	0 to 1	73	$ 66,000	$ 69,345	$ 77,500
		2	35	$ 65,000	$ 67,000	$ 69,918
		3	35	$ 64,000	$ 67,700	$ 72,000
		4 to 5	47	$ 66,000	$ 69,900	$ 75,000
		6 or more	12	$ 63,900	$ 67,750	$ 69,910
	Associate Professor	0 to 1	22	$ 71,900	$ 82,275	$ 95,134
		2 to 3	43	$ 80,000	$ 85,100	$ 89,000
		4 to 5	31	$ 70,000	$ 87,700	$ 91,900
		6 to 8	30	$ 68,538	$ 78,021	$ 82,441

From Amstat News

IV. Transitions: Research/Job

Industry

Years of Experience	Highest Degree	N	Q1	Median	Q3
With No Managerial Responsibility					
0–1.9	BS MS PhD	3 25 18	 55.0 72.0	 **60.0** **83.0**	 70.0 85.0
2–3.9	BS MS PhD	3 28 22	 64.0 85.0	 **70.0** **88.0**	 80.0 98.0
4–7.9	BS MS PhD	4 62 71	 67.0 85.0	 **75.0** **94.0**	 85.0 104.0
8–11.9	BS MS PhD	2 45 36	 70.0 88.0	 **83.0** **98.5**	 94.0 121.0

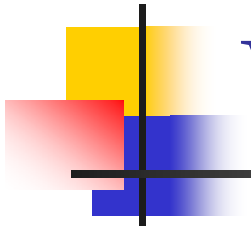
From Amstat News

IV. Transitions: Research/Job

Government

Years of Experience	Highest Degree	N	Q1	Median	Q3
With No Managerial Responsibility					
0–1.9	MS PhD	1 1			
2–3.9	BS MS PhD	2 3 4			
4–7.9	MS PhD	8 18	82.0	**78.0** **85.0**	94.0
8–11.9	MS PhD	5 11	79.0	**85.0**	98.0

From Amstat News



V. Some final words: course preparation

Life Sciences

- Cell Biology/Molecular Biology
- Biochemistry
- Genetics

Computer Science

- Intermediate/Advanced Programming (JAVA, C++)
- Fundamental Data Structures and Algorithms
- Algorithms

Physical Sciences

- Statistical Thermodynamics or Physical Chemistry

Mathematics and Statistics

- Vector Calculus
- Linear Algebra
- Probability & Statistics

Computational Biology

- Computational Biology; Bioinformatics

V. Some final words: Taiwan or abroad

- Try to go abroad if possible

- There are very good graduate programs in Taiwan. If you choose to stay, try to apply for a one-year exchange program abroad.

V. Some final words: Preparation

- Course preparation
- Improve English (take GRE and TOEFL early)
- Talk to some researchers in NTU and Sinica
- Get good recommendation letters and write a good essay
- Go to talks (NTU Math, NTU biostatistics, Sinica)
- Apply as many (good) schools as you can.
- Money should not be an issue if you get stipend support.

V. Some final words: after you get there

- Continue to improve English
- Find a good advisor (reputation in research, personality)
- Be collegial and collaborative; change our viewpoint and re-interpret what you see without bias.

Thanks for your attention!

Merry Christmas!!