Selected Topics in Bioinformatics

Hao Bailin[a,b,c]

a. T-Life Research Center
Fudan University, Shanghai 200433

b. Beijing Genomics Institute (BGI)
and Hangzhou Branch of BGI

c. Institute of Theoretical Physics
Academia Sinica, Beijing 100080

http://www.itp.ac.cn/~hao/

An Approximate Plan of Lectures

1. Physics and Biology
2. Brief introduction to molecular biology
3. Biological data and challenge of big numbers
4. Computer and computer science: a summary
5. *Addendum*: How to extend one's English vocabulary
6. Brief introduction to probability, statistics and statistical physics
7. Sequence models
8. Gene-finding in genomes
9. Language and combinatorics: avoidance pattern in prokaryote genomes
10. Fine structure in 1D histograms of some randomized genomic sequences

11. Graph theory: decomposition and reconstruction of protein sequences
12. Multi-alignment and phylogenetic trees
13. Case study: prokaryote phylogeny without sequence alignment

• **Bioinformatics** — coined by H. A. Lim in early 1990s, although the term appeared in a book title in a different context in 1989.

Data-driven study, knowledge discovery from data, data-mining

In a narrow sense: sequence analysis

Example: Given a newly sequenced sequence, search for homologs in all known DNA databases.

• **Biocomputing** — more knowledge-based study

• **Computational biology**

Never start from a formal definition only.

Example: Prediction of 3D structure of
a protein

By *ab initio* molecular dynamics calculation
→ computational biology.

By *threading* → bioinformatics.

A Few References

1. A. D. Baxevanis, B. F. F. Ouellette, eds. *Bioinformatics. A practical Guide to the Analysis of Genes and Proteins*, Wiley-Interscience, 1998.

2. T. K. Attwood, D. J. Parry-Smith, *Introduction to Bioinformatics*, AWL Press, 1999.

3. D. W. Mount, *Bioinformatics: Sequence and Genome Analysis*, Cold Spring Harbor Laboratory Press, 2000.

4. Hao Bailin, Zhang Shuyu, *Handbook of Bioinformatics*, Shanghai SciTech Press (in Chinese), 1st Ed. 2000; 2nd Ed. 2002.

Some Useful URLs

1. National Center for Biotechnological Information (NCBI)
   `http://www.ncbi.nlm.nih.gov`
2. European Bioinformatics Institute (EBI)
   `http://www.ebi.ac.uk`
3. DNA Data Bank of Japan
   `http:www.ddbj.nig.ac.jp`
4. Genomes of Human and other model organisms
   `http://www.ensembl.org/`
5. Expert Protein Analysis System (ExPASy) including SWISS-PROT
   `http://www.expasy.ch/`
6. Center for Bioinformatics, Peking University
   `http://www.cbi.pku.edu.cn/`
7. Center for Bioinformatics, Shanghai Institutes of Life Science, Academia Sinica
   `http://www.BioSino.org`
8. "Google University"
   `http://google.com`

Bioinformatics Challenge

BI challenge is the challenge of big numbers.

- **Numbers in Macro-Biology**:

Number of species on Earth: estimation from 2 Mil to 2 Bil

More than 154 000 species represented by at least one sequence in GenBank (Nov. 2003)

Only about 50 000 have some taxonomic information.

About 5000 mammalian species exist out of 200 000 that ever existed.

Bacteria — by far the most successful species (procaryote). How many there are?

- **Numbers related to Molecular Biology**.

Exponential Growth of GenBank Data

Rel.	Date	Seq(10^6)	bp(10^9)	Aver.
104	15 Dec. 97	1.25	1.891	665
110	5 Dec. 98	3.04	2.162	710
115	15 Dec. 99	5.35	4.654	869
121	15 Dec. 00	10.09	11.096	1099
127	15 Dec. 01	14.97	15.849	1058
128	15 Jan. 02	15.47	17.089	1105
129	15 Apr. 02	16.76	19.073	1137
130	15 Jun. 02	17.47	20.649	1182
131	15 Aug. 02	18.19	22.616	1242
132	15 Oct. 02	19.80	26.525	1339
133	15 Dec. 02	22.31	28.507	1277
134	15 Feb. 03	23.03	29.358	1274
135	15 Apr. 03	24.02	31.099	1294
136	15 Jun. 03	25.59	32.528	1271
137	15 Aug. 03	27.21	33.865	1244
138	15 Oct. 03	29.81	35.599	1193
139	15 Dec. 03	30.96	36.553	1180
140	15 Feb. 04	32.54	37.893	1164
141	15 Apr. 04	33.67	38.989	1157

A few Big Numbers

Daily production of sequencing data at BGI: 3×10^7 (10^{10} yearly)

Rice Genome *indica*: 4.3×10^8 bp

Human Genome size: 3.2×10^9 bp

Biodata produced yearly worldwide at present: 10^{15} bytes

Yearly increase of hard disks at the Sanger Center: 100TB=10^{14} Bytes

• Time elapsed since the Big Bang: 4×10^{17} seconds

• Words ever spoken by all mankind: $\sim 10^{18}$

Units of Big and Small Numbers

Unit	Name	Unit	Name
10^{3}	Kilo	10^{-3}	milli
10^{6}	Mega	10^{-6}	micro
10^{9}	Giga	10^{-9}	nano
10^{12}	Tera	10^{-12}	pico
10^{15}	Peta	10^{-15}	femto
10^{18}	Exa	10^{-18}	atto

Some Formulae Related to Big Numbers

$\left(1+\frac{x}{N}\right)^N \to e^x \quad N \gg 1$

$\left(1-\frac{x}{N}\right)^N \to e^{-x} \quad N \gg 1$

$n!$ grows fast, but not as fast as n^n.

Stirling's formula:

$n! \approx \sqrt{2\pi n}\ \frac{n^n}{e^n}$

Good enough for any practical purpose:

n	$n!$	Stirling	Error
10	3.63×10^6	3.60×10^6	8.3×10^{-3}
18	6.402×10^{15}	6.373×10^{15}	4.6×10^{-3}

Binomial Explosion

Binomial coefficients: number of combinations of picking up y objects from a total of n objects:

$$C_n^y \equiv \binom{n}{y} = \frac{n!}{y!(n-y)!}$$

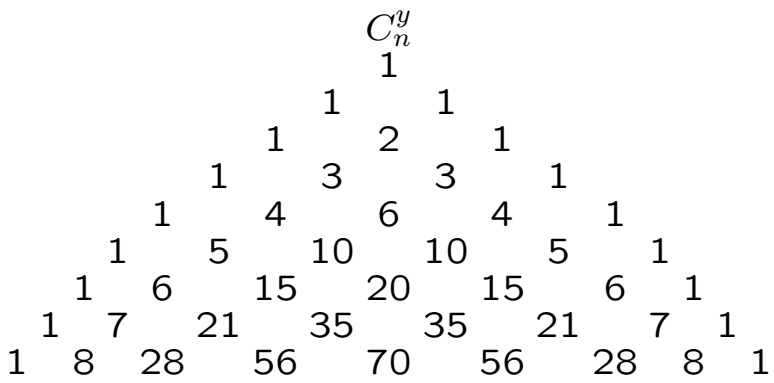
Its central members (for $y \approx n/2$) grow fast but not as fast as $n!$.

Using Stirling's formula:

$$C_n^y \approx \frac{1}{\sqrt{2\pi}} \frac{\sqrt{n}}{\sqrt{y(n-y)}} \left(\frac{y}{n}\right)^{-y} \left(1 - \frac{y}{n}\right)^{-(n-y)}$$

Limitation of exhaustive enumeration algorithms.

The Yang Hui Triangle



Computational Complexity

Space-wise (memory) and time-wise (CPU).

• N: scale of the problem (number of letters in a sequence, size of a matrix, etc.)

• Growth of computing time with N:

1. $\propto \log N$
2. $\propto N$ (linear)
3. $\propto N^2$, or, in general, $\propto$ a polynomial of N
4. $\propto e^N$, or $N!$, or $N^N, \cdots$. Impossible to treat. (NP-hard and NP-complete)

• Growth of memory size: likewise.

• Interchange between time and space.

"A constant space linear-time algorithm"