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
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


Some Useful URLs

1. National Center for Biotechnological Information (NCBI)
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
8. “Google University”
   http://google.com
Biinformatics 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

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

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

Rice Genome *indica*: \(4.3 \times 10^8\) bp

Human Genome size: \(3.2 \times 10^9\) bp

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

Yearly increase of hard disks at the Sanger Center: \(100\,\text{TB}=10^{14}\) Bytes

- Time elapsed since the Big Bang: \(4 \times 10^{17}\) seconds

- Words ever spoken by all mankind: \(~10^{18}\)
## Units of Big and Small Numbers

<table>
<thead>
<tr>
<th>Unit</th>
<th>Name</th>
<th>Unit</th>
<th>Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>$10^3$</td>
<td>Kilo</td>
<td>$10^{-3}$</td>
<td>milli</td>
</tr>
<tr>
<td>$10^6$</td>
<td>Mega</td>
<td>$10^{-6}$</td>
<td>micro</td>
</tr>
<tr>
<td>$10^9$</td>
<td>Giga</td>
<td>$10^{-9}$</td>
<td>nano</td>
</tr>
<tr>
<td>$10^{12}$</td>
<td>Tera</td>
<td>$10^{-12}$</td>
<td>pico</td>
</tr>
<tr>
<td>$10^{15}$</td>
<td>Peta</td>
<td>$10^{-15}$</td>
<td>femto</td>
</tr>
<tr>
<td>$10^{18}$</td>
<td>Exa</td>
<td>$10^{-18}$</td>
<td>atto</td>
</tr>
</tbody>
</table>
Some Formulae Related to Big Numbers

\[(1 + \frac{x}{N})^N \rightarrow e^x \quad N \gg 1\]

\[(1 - \frac{x}{N})^N \rightarrow e^{-x} \quad N \gg 1\]


\[n! \text{ grows fast, but not as fast as } n^n.\]

**Stirling’s formula:**

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

Good enough for any practical purpose:

<table>
<thead>
<tr>
<th>(n)</th>
<th>(n!)</th>
<th>Stirling</th>
<th>Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>(3.63 \times 10^6)</td>
<td>(3.60 \times 10^6)</td>
<td>(8.3 \times 10^{-3})</td>
</tr>
<tr>
<td>18</td>
<td>(6.402 \times 10^{15})</td>
<td>(6.373 \times 10^{15})</td>
<td>(4.6 \times 10^{-3})</td>
</tr>
</tbody>
</table>
**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

\[
\begin{array}{cccccccc}
C_n^y \\
1 \\
1 & 1 \\
1 & 3 & 3 & 1 \\
1 & 4 & 6 & 4 & 1 \\
1 & 5 & 10 & 10 & 5 & 1 \\
1 & 6 & 15 & 20 & 15 & 6 & 1 \\
1 & 7 & 21 & 35 & 35 & 21 & 7 & 1 \\
1 & 8 & 28 & 56 & 70 & 56 & 28 & 8 & 1 \\
\end{array}
\]
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 $N^N$, $\ldots$ Impossible to treat. (NP-hard and NP-complete)

- Growth of memory size: likewise.

- Interchange between time and space.

"A constant space linear-time algorithm"