Chinese

« Bombardment of the Chinese Embassy in Belgrade »

Sideways

Context-driven

Anglo-American	Greco-Latin
NATO	OTAN
Bottom Up	Top Down
Data-driven	Hypothesis-driven

What is Life?

- Physics: *matter, energy, time*

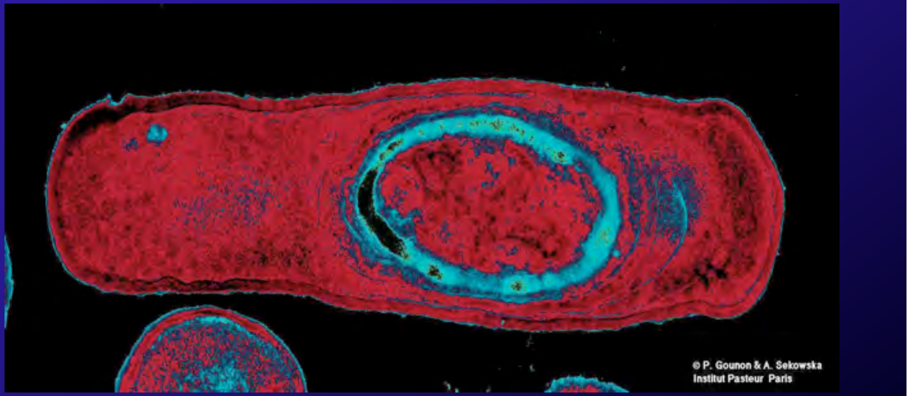
- Biology: Physics + *information, coding, control...*

What is Life?

Three processes are needed for Life:

- Metabolism ("eating and digesting")
- Compartmentalization (an "inside" and an "outside")
- Information transfer (a "program")

The atom of life is the Cell

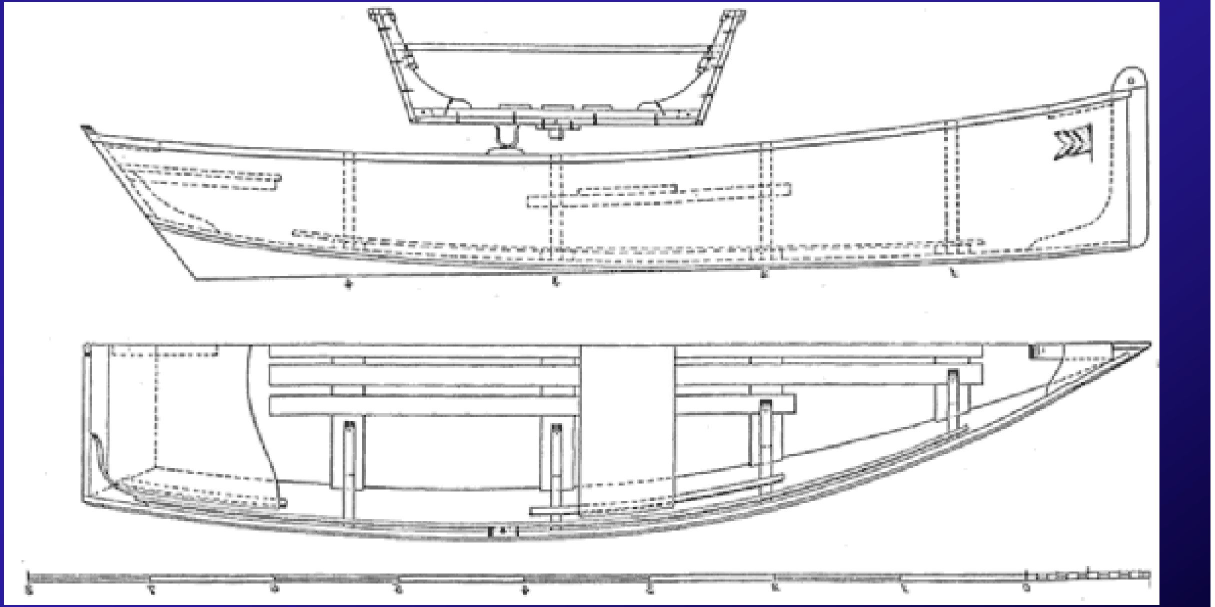


© P. Gounon & A. Sekowska
Institut Pasteur Paris

香港大學－巴斯德研究中心
HKU-Pasteur Research Centre

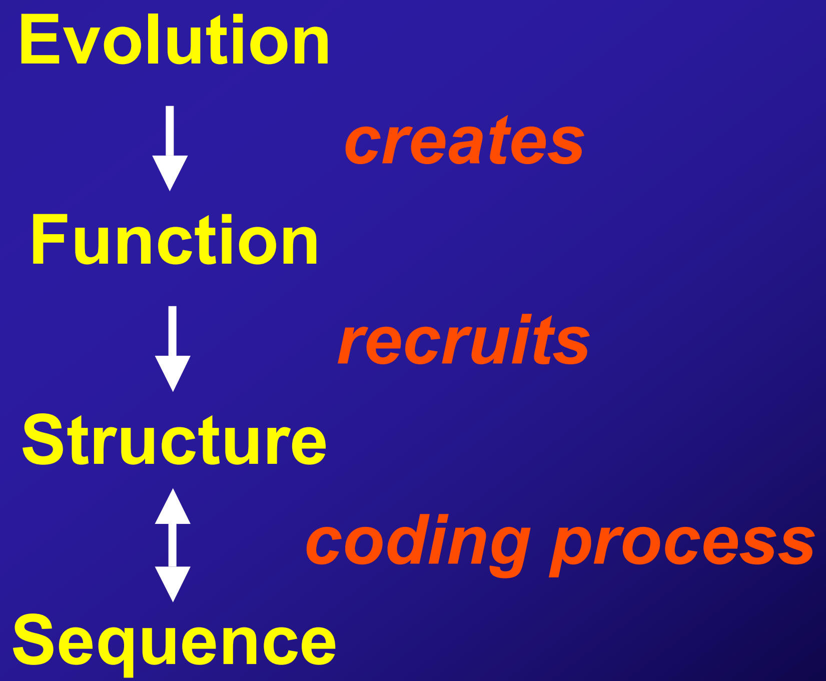
The Delphic Boat

- Genes do not operate in isolation
- Proteins are part of complexes, as are parts in an engine
- ➔ It is important to understand their relationships, as those in the planks which make a boat



Empedocles / Maupertuis / Malthus / Darwin

Variation / Selection / Amplification



Evolution

creates

Function

recruits

Structure

coding process

Sequence

Different levels of information

- What is seen by replication: no meaning, Shannon ’s information
- What is seen by the gene expression machinery
  - Algorithmic complexity (space)
  - Logical depth (time)
  - (Critical depth) (finiteness)

Shannon's entropy (1)

Caveat: Myron Tribus relates that von Neumann, to whom Shannon had turned to help him find a name for his function defining information, proposed prophetically: "*You should call it entropy for two reasons. In the first place your uncertainty function has been used in statistical mechanics under that name, so it already has a name. In the second place, and more important, no one knows what entropy really is, so in a debate you will always have the advantage*", thus opening a Pandora's box of intellectual confusion.

Shannon's entropy (2)

$$H(p_i) = -\Sigma\,\{p_i \log_2 p_i \mid i \in I\}$$

Note that the validity of this formula rests on very strong hypotheses about the nature of the signals (in particular that the signals fit standard Laplace-Gauss probability laws)

Shannon's entropy (3)

- What is seen by replication: signals can be identified because they are information poor (in the sense of Shannon) when they form a « consensus »
- Note that the genetic code would not belong to this category, since it is possible to find out (autocorrelation) the existence of a period of three with no consensus...

Principal Component Analysis

Giving a set of multivariate measurements the purpose is to find a smaller set of variables with less redundancy, while preserving the quality of the data set.

Centered normalized measurements are used, and an orthogonal coordinate system is identified in which the redundancy induced by correlations has disappeared. The variance of the projections of the data on the new coordinate systems is also maximized.

Factorial Correspondence Analysis

This is a type of PCA where the data are not simply centered and normalized but measured by their distances using the chi-square test.

This type of analysis gives less weight to isolated values and to small sets with particular properties.

Clustering Method: Dynamic Clouds (K-means)

Starting with an arbitrary number of classes, with a seed, one computes a partition, using each seed as a barycentre for a given measure

Using this partition, a barycentre is computed, and used as a seed to compute a new partition

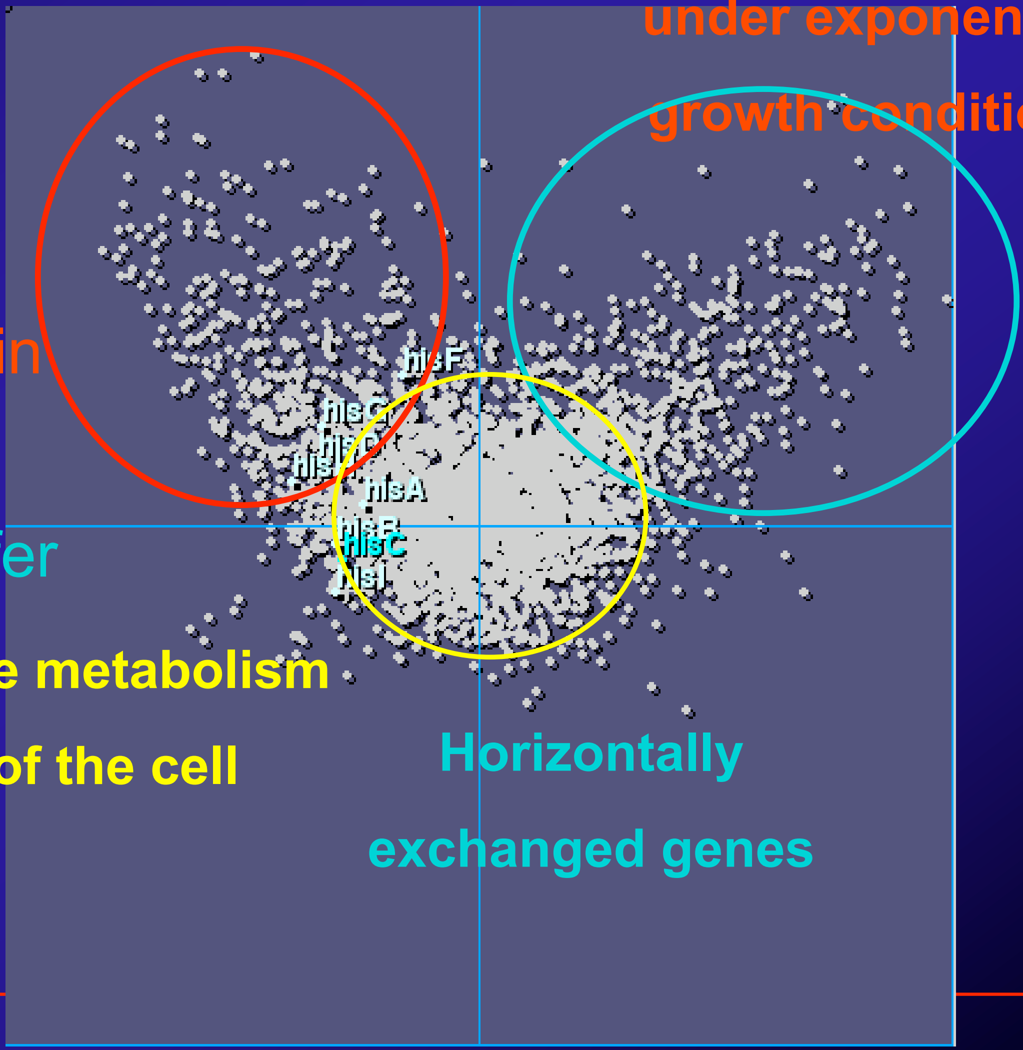
Etc.

Neighborhoods

Class I: core metabolism

Class II: high expression in exponential growth

Class III: horizontal transfer



Genes expressed at a high level under exponential growth conditions

Core metabolism of the cell

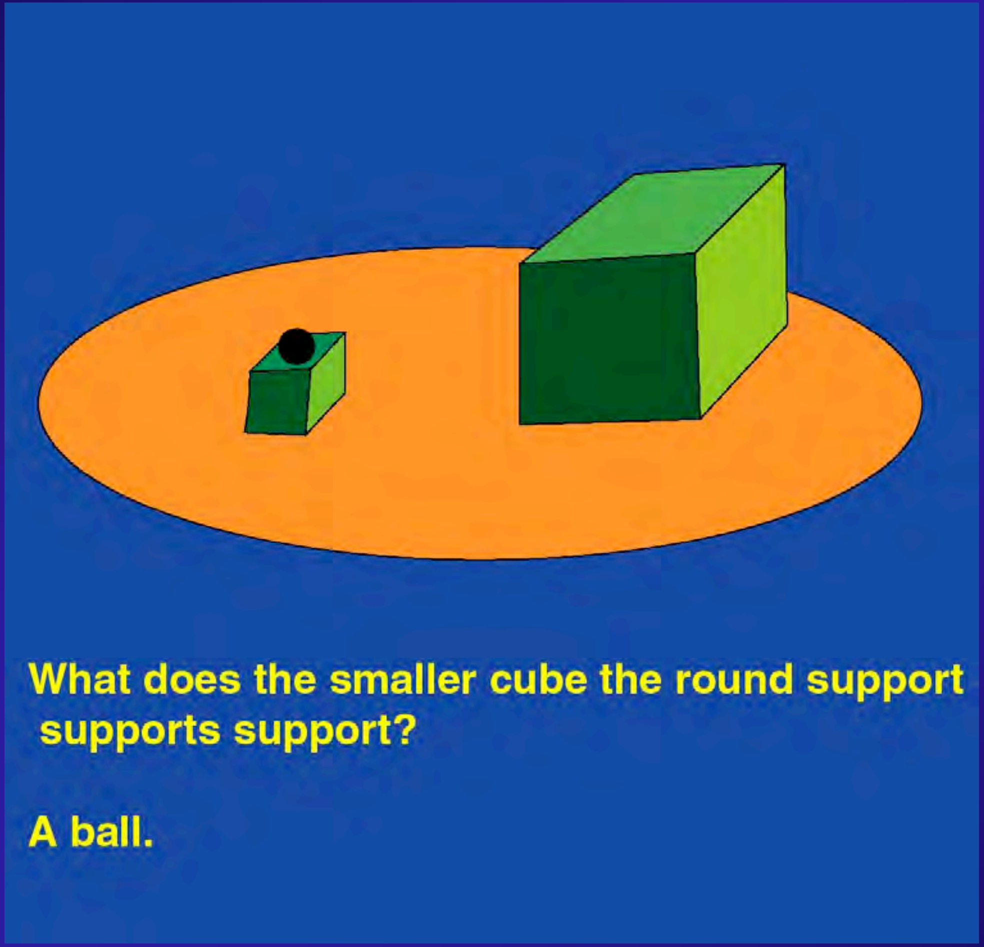
Horizontally exchanged genes

Algorithmic complexity

- Shannon's entropy works on collections of messages
- Kolmogorov and others proposed to define randomness of one sequence by stating that it cannot be described by a program with a length shorter than the sequence

- This provides us a **research program**: in order to approach algorithmic complexity of a sequence, we need to describe how it has been constructed (in the real physical world)
- **Prokaryotes look « random »; eukaryotes look « repeated »**

Repeats



What does the smaller cube the round support supports support?

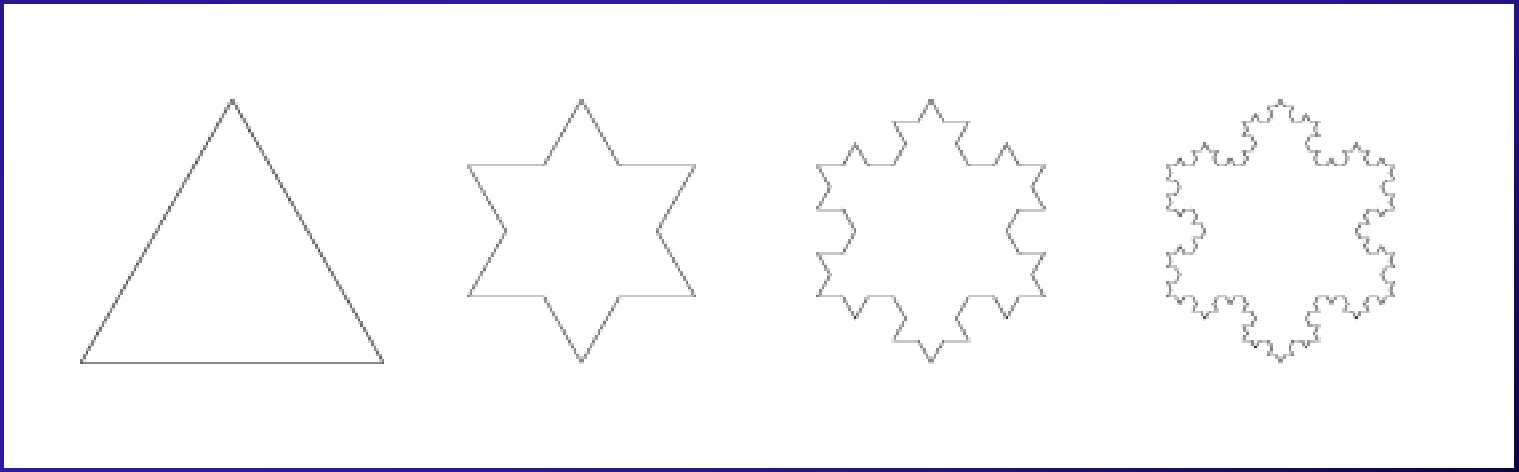
A ball.

Remember also:

This clock has a minute minute hand

Logical Depth (1)

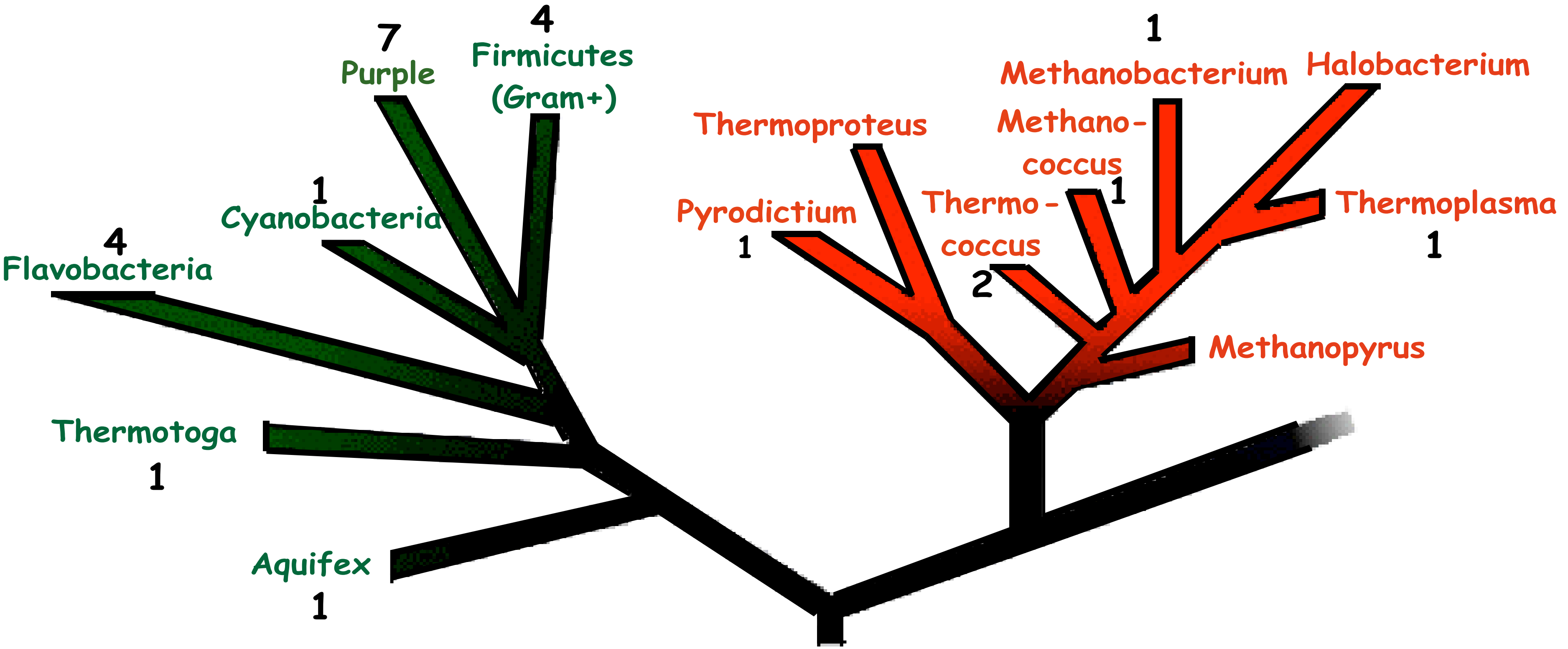
A very short program can describe a repeated sequence, or a fractal figure such as Koch 's snowflake



Logical Depth (2)

- The difference is on the value of the information provided on the nth step.
- Recursive programs imply that it is necessary to run the program to get the information
- Bennett named « logical depth » the time needed to get the information

Evolution



(Woese, 1990)

Mutations

```
 1    TTAAG TGA GGGCGAAAAGAAACT ATG GAT AAA TGG CTC ATG CAA TAT AAA TTA --- GCT AGA GAA GAG CTT TCT AAA
 7    TTAAG TGA GGGCGAAAAGAAACT ATG GAT AAA TGG CTC ATG CAA TAT AAA TTA --- GCT AGA GAA GAG CTT TCT AAA
15    TTAAG TGA GGGCGAAAAGAAACT ATG GAT AAA TGG CTC ATG CAA TAT AAA TTA --- GCT AGA GAA GAG CTT TCT AAA
18    TTAAG TGA GGGCAAAAAGAAACT ATG GAT GAA TGG CTC ATG CAA TAT AAA TTA --- GCT AGA GAA GAG CTT TCT AAA
28    TTAAG TGA GGGCGAAAAGAAACT ATG GAT AAA TGG CTC ATG CAA TAT AAA TTA --- GCT AGA GAA GAG CTT TCT AAA
30    TTAAG TGA GGGCGAAAAGAAACT ATG GAT AAA TGG CTC ATG CAA TAT AAA TTA --- GCT AGA GAA GAG CTT TCT AAA
44    TTAAG TGA GGGCGAAAAGAAACT ATG GAT AAA TGG CTC ATG CAA TAT AAA TTA --- GCT AGA GAA GAG CTT TCT AAA
26695 ttaag tga gggcgaaaaggaact atg gat aaa tgg ctc atg caa tat aaa ttg --- gct aga gaa gag ctt tct aaa
J99   ttaag tga gggcaacaagagact atg gat aaa tgg ctc atg caa tac aga ttg --- gct aga gaa gag ctt tct aaa
                    RBS
HP                               M   D  K(E) W   L   M   Q   Y  K(R) L   -   A   R   E   E   L   S   K
CJ                               M   E   K   L   I   T   Y   F   K   L   -   S   K   A   E   L   R   K
DR                               M   N   -   L   I   Q   Y   F   R   D   -   A   R   E   E   L   S   R
EC   Two transmembrane domains   G   K   A   T   V   A   F   A   R   E   -   A   R   T   E   V   R   K
BS                               M   R   -   I   M   K   F   F   K   D   V   G   K   -   E   M   K   K
```

Mutual information

With standard metrics:

$$H(p_{IJ}\,;p_I\,.\,p_J\,) = H(p_I) + H(p_J\,) — H(p_{IJ})$$

$$= \Sigma\{p_{ij}\log_2(p_{ij}\,/p_i\,.\,p_j)\mid i\in\ I, j\in\ J\}$$

$$= \Sigma\{p_i\,.\,p_j(p_{ij}\,/p_i\,.\,p_j)\log_2(p_{ij}\,/p_i\,.\,p_j)\mid i\in\ I, j\in\ J\}$$

$$= \Sigma\ \{p_i\,.\,p_j\ f(p_{ij}\,/p_i\,.\,p_j)\mid i\in\ \ I, j\in\ J\}$$